Machine Learning for the Prediction of Protein-Protein Interactions

by

José Antonio Reyes

A dissertation submitted to

The Department of Computing Science

of

The University of Glasgow

for the degree of

Doctor of Philosophy

November, 2009

Abstract

The prediction of protein-protein interactions (PPI) has recently emerged as an important problem in the fields of bioinformatics and systems biology, due to the fact that most essential cellular processes are mediated by these kinds of interactions. In this thesis we focussed in the prediction of co-complex interactions, where the objective is to identify and characterize protein pairs which are members of the same protein complex.

Although high-throughput methods for the direct identification of PPI have been developed in the last years. It has been demonstrated that the data obtained by these methods is often incomplete and suffers from high false-positive and false-negative rates. In order to deal with this technology-driven problem, several machine learning techniques have been employed in the past to improve the accuracy and trustability of predicted protein interacting pairs, demonstrating that the combined use of direct and indirect biological insights can improve the quality of predictive PPI models. This task has been commonly viewed as a binary classification problem. However, the nature of the data creates two major problems. Firstly, the imbalanced class problem due to the number of positive examples (pairs of proteins which really interact) being much smaller than the number of negative ones. Secondly, the selection of negative examples is based on some unreliable assumptions which could introduce some bias in the classification results.

The first part of this dissertation addresses these drawbacks by exploring the use of one-class classification (OCC) methods to deal with the task of prediction of PPI. OCC methods utilize examples of just one class to generate a predictive model which is consequently independent of the kind of negative examples selected; additionally these approaches are known to cope with imbalanced class problems. We designed and carried out a performance evaluation study of several OCC methods for this task. We also undertook a comparative performance evaluation with several conventional learning techniques.

Furthermore, we pay attention to a new potential drawback which appears to affect the performance of PPI prediction. This is associated with the composition of the positive gold standard set, which contain a high proportion of examples associated with interactions of ribosomal proteins. We demonstrate that this situation indeed biases the classification task, resulting in an over-optimistic performance result. The prediction of non-ribosomal PPI is a much more difficult task. We investigate some strategies in order to improve the performance of this subtask, integrating new kinds of data as well as combining diverse classification models generated from different sets of data.

In this thesis, we undertook a preliminary validation study of the new PPI predicted by using OCC methods. To achieve this, we focus in three main aspects: look for biological evidence in the literature that support the new predictions; the analysis of predicted PPI networks properties; and the identification

of highly interconnected groups of proteins which can be associated with new protein complexes.

Finally, this thesis explores a slightly different area, related to the prediction of PPI types. This is associated with the classification of PPI structures (complexes) contained in the Protein Data Bank (PDB) data base according to its function and binding affinity. Considering the relatively reduced number of crystalized protein complexes available, it is not possible at the moment to link these results with the ones obtained previously for the prediction of PPI complexes. However, this could be possible in the near future when more PPI structures will be available.

Acknowledgements

First, I would like to thank my supervisor Professor David Gilbert for all the support, encouragement and advice I received from him during the time I have been doing my research.

Many thanks go to my friends and to all members, past and present, of the Bioinformatics Research Centre, who share some time, conversations and experiences all these years I have spent in Glasgow.

I am very grateful to my parents and family for their love, constant support and encouragement to pursue this challenge. Without them this thesis would not have been possible. Thanks for always be there when I needed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The prediction of protein-protein interactions (PPI) has recently emerged as an important problem in the fields of bioinformatics and systems biology, due to the fact that most essential cellular processes are mediated by these kind of interactions. These processes include cell cycle control, differentiation, signalling, transcription and transport.

Traditionally PPI have been identified through the use of small scale experimental techniques, which allow the correct and accurate identification of these kind of interactions. However these small scale methods are expensive and time consuming. Thus it is not feasible to investigate all potential protein pairs in this way. In fact currently most PPI remain undiscovered (von Mering et al. 2002, Futschik et al. 2007).

High-throughput methods for the direct identification of PPI have been recently developed including yeast two-hybrid screens (Y2H) (Uetz et al. 2000, Ito et al. 2001) for detection of binary physical interactions and mass spectrometry methods for protein complex identification (Gavin et al. 2002, Ho et al. 2002).

Even though high-throughput techniques can increase the number of predicted PPI, in general the data obtained by these methods is often incomplete and suffers from high false-positive and false-negative rates (von Mering et al. 2002).

Considering the limitations of the available experimental techniques for the detection of new PPI, additional approaches are needed in order to improve the accuracy and trustability of predicted protein interacting pairs. Various studies have previously been developed, based on the integrative learning analysis of diverse biological sources of information (Bader et al. 2004, Lanckriet et al. 2004, Gilchrist et al. 2004, Lee et al. 2004, Yamanishi et al. 2004). These have demonstrated that the combined use of direct and indirect biological insights can improve the quality of predictive PPI models.

The prediction of PPI has been commonly viewed as a classical binary classification problem, where the aim is to predict whether any two proteins do or do not interact. Several traditional machine learning methods have been employed in the past for this specific task (Jansen et al. 2003, Lin et al. 2004, Zhang et al. 2004, Lu et al. 2005, Qi et al. 2005, Ben-Hur and Noble 2005, Qi et al. 2006). These methods generally use supervised learning algorithms where the final objective is to generate a classification model from a gold standard reference set of positive (truly interacting protein pairs) and negative examples (non-interacting pairs). Two main drawbacks have been identified regarding these previous approaches:

i) in general they face a highly imbalanced classification problem, where the number of positive examples is much smaller than the number of negative examples. This affects the quality of the predictive models which may be biased towards the majority class and consequently the minority class examples are poorly predicted. Under-sampling and cost-sensitive strategies have been used to deal with the imbalanced problem in some of these previous works, whilst

others did not report any action taken to deal with.

ii) Although the selection of positive examples is based on reliable experimental techniques (i.e. small scale experiments), there is no experimental method to find pairs of proteins which do not interact (negative examples). Therefore, certain assumptions have to be made in order to construct a negative gold standard set, which can introduce some bias into the learning process and consequently produces a significant effect upon the performance of the classification approach (Ben-Hur and Noble 2006).

In order to deal with this situation, we propose the use of one-class classification (OCC) methods in this research as a possible solution to these issues. The aim of OCC is to use feature information from only one of the classes, positive examples in this case, to generate a classification model. OCC methods are known to be able to deal efficiently with highly imbalanced classification problems (Chawla et al. 2004). Additionally, unlike conventional binary classifiers, OCC methods produce classification models which are independent of the kind of negative gold standard set employed. In this thesis we designed and carried out a performance evaluation study of several OCC methods for this task. Among them we have found that the Parzen density estimation approach outperforms the others. We also undertook a comparative performance evaluation between the Parzen OCC method and several conventional learning techniques, considering different scenarios, for example varying the number of negative examples used for training purposes. We found that the Parzen OCC method in general performs competitively with traditional approaches and in many situations outperforms them.

Another potential drawback associated with this prediction task derives from the composition of the positive gold standard set. Some of the protein complexes included in the reference data set are bigger than the average. For instance, the

ribosomal complexes (large and small subunits) represent almost 2/3 of the total of PPI available in the positive reference data set commonly employed for yeast. This situation could have an important effect on the performance of the different classification techniques employed. This situation has not been investigated before in related literature. In this thesis, we intend to address this potential problem, demonstrating that the high proportion of ribosomal-protein based examples does indeed create bias in the PPI prediction results. When removing these ribosomal-based PPI, we face a more difficult prediction task. In order to improve the performance of this subtask we first integrate new biological features based on protein secondary structure information. Subsequently we investigated and demonstrated that, by combining the predictions of several Parzen OCC models induced from different subsets of biological data, it is possible to significantly increment the performance of this subtask.

The final goal associated with the use of computational methods for predicting PPI is to predict or identify new potential PPI targets. These potential targets can then be used, for instance, to guide biologists developing small scale experiments in order to validate them. In this thesis we undertake a preliminary evaluation analysis of the capability of the Parzen OCC approach to predict new potential PPI targets. For this we generate a set of PPI consisting of random protein pairs not employed to generate the predictive model. We then apply the parzen OCC model to this random set to predict new potential PPI among them. In this thesis we focus our analysis in three main topics: firstly, we analyze the topology properties of the PPI network predicted; secondly we look for highly interconnected groups of proteins which can be associated to new protein complexes; finally, we look for evidence in the related biological literature and data bases to validate these new predictions.

The final part of this thesis explores a slightly different area, related to the

prediction of PPI types. This is associated with the classification of PPI structures (complexes) contained in the Protein Data Bank (PDB) data base according to its function and binding affinity. Protein structures are obtained by experimental techniques such as X-Ray crystallography. These structures can be classified according to their life time and binding affinity into four main classes, as obligate permanent interactions involving homo or hetero obligomers (Nooren and Thornton 2003) and non-obligate transient interactions involving Enzyme-inhibitor or non Enzyme-inhibitor (Bradford and Westhead 2004). Here we introduce a novel computational approach for the prediction of PPI types employing association rule based classification (ARBC). This includes association rule generation and posterior classification based on the discovered rules. We investigate diverse properties associated with the interaction interface of crystalized protein complexes, aiming to discover patterns in the form of association rules that correctly classify PPI types and at the same time characterize PPI binding sites. Due the complexity of experimental techniques, at present there is a reduced number of available protein complexes structures. Consequently there is not enough examples available to link these results with the ones obtained previously for the prediction of PPI complexes. However in the near future is expected that the number of crystalized protein structures will be increased. In this case the information related to PPI types can be useful to enhance the predictions made by our previous techniques (one-class classification).

1.2 Contributions

Here we present the main contributions to knowledge associated to this thesis:

- We investigated the use of OCC models for the task of PPI prediction. A comparative performance evaluation between OCC and conventional clas-

sifiers for the task of PPI prediction was developed. It was demonstrated that PPI can be predicted effectively using OCC methods, especially considering the Parzen OCC approach.

- We demonstrated that OCC models deal efficiently with the imbalanced class problem associated to this task. On the contrary conventional classifiers are strongly affected by this situation.

- We investigated the problem of reliability of negative gold standard in the prediction of PPI task. It was demonstrated that the performance of conventional classifiers is highly affected by the quantity of negative data employed on either training and testing the models.

- A new drawback not reported previously in literature for the prediction of PPI was investigated. This problem is related to the high proportion of ribosomal-based proteins in the positive gold standard reference data set generally employed for this task. We demonstrated that this situation is biasing the results of classifiers and consequently the reliability of new predicted PPI. The task associated to the prediction of non-ribosomal PPI is much more difficult.

- Focused in the sub-task of prediction of non-ribosomal PPI, we investigated the use of protein secondary structure (SS) information into this problem. Features based on this kind of information have not been employed before for this task. We demonstrated that protein SS features have a positive effect, improving significantly the performance of OCC and conventional classifiers on this task.

- Following with the task of prediction of non-ribosomal PPI. We investigated several strategies for the combination of various Parzen OCC models generated from diverse sets of biological information. We demonstrated

that a the combination approach improves significantly the performance of this subtask.

- To select a set of appropriate OCC models for combination purposes, we investigated the diversity of their predictions. In this thesis we adapted several diversity measures focusing in the low false-positive region of different classification models evaluated.

- We developed a preliminary validation study of new PPI predicted employing OCC models. We demonstrated that the PPI network associated to these predictions, share similar topological properties with other PPI and biological networks previously reported in the related literature. We also identify several PPI clusters (highly connected proteins), which can be associated for instance to new protein complexes. We demonstrated that it is possible to infer new biological knowledge from the analysis of these clusters.

- We implemented a novel approach for prediction of PPI types using association rule based classification (ARBC) approach. We demonstrated that ARBC performs competitively with other classifiers, but additionally with the advantage of improving the interpretability of the predictive results. The main contribution of this thesis was related to the implementation of the classification stage.

1.3 List of Publications

The list of publications associated to this thesis is as follows:

- Jose A. Reyes and David Gilbert. Prediction of protein-protein interactions using one-class classification methods and integrating diverse biological

data. Journal of Integrative Bioinformatics, 4(3):77, 2007. ISSN 1613-4516

- Jose A. Reyes and David Gilbert. Combining One-Class Classification Models Based on Diverse Biological Data for Prediction of Protein-Protein Interactions. In Proceeding of Data Integration and the Life Sciences DILS 2008. LNCS/LNBI 5109, pp 177-191, 2008. Springer-Verlag .

- Sung Hee Park, Jose A. Reyes, David Gilbert, Sang Soo Kim and Ji Woong Kim. Prediction of Protein-Protein Interaction Types using Association Rule based Classification. BMC Bioinformatics 2009, 10(36).

1.4 Thesis Organization

Chapter 2 provides a brief introduction to biological background relevant to understand further parts of this thesis. Chapter 3 present an overview of machine learning techniques employed in the past to deal with the problem of prediction of PPI. Chapter 4 exhibits the work associated to the use of OCC models for the problem of PPI prediction. A comparative performance evaluation with conventional classification methods is also carried out. In Chapter 5, we investigated several strategies to improve the performance of OCC models for this task. Including, the integration of new biological features, and the combination of several OCC models based on diverse sets of biological information. Chapter 6 presents a preliminary validation study of the new PPI predicted employing OCC models. In Chapter 7, we describe a computational approach for the prediction of PPI types employing association rule based classification (ARBC). Finally, Chapter 8 presents the conclusions of this thesis and ideas for future work.

Chapter 2

Biological Background

2.1 Introduction

In this chapter, we present an overview of the basic concepts of molecular biology relevant to understanding further parts of this thesis, including information about DNA, proteins, protein-protein interactions (PPI) and protein interaction networks.

In this thesis, we refer to the term "protein-protein interaction" as the association of two proteins with each other. PPI can be classified according to different properties which will be examined here. We specifically focus on interactions related to protein complexes. In this case, any two proteins interact with each other if they are members of the same complex. We will briefly describe the main experimental techniques available today to identify these kinds of biological interactions, and we will discuss their capabilities and limitations. Finally, we will introduce the major existing biological databases related to this kind of information.

2.2 Proteins

Proteins are essential macromolecules which are involved in almost all processes in the cell. They are fundamental structural components of cells and are also involved in almost every cell function such as transportation, hormonal regulation, metabolism, respiration, repair and control of genes. Proteins do not usually work alone but interact with other proteins, forming protein complexes and also protein interaction networks. For this reason, understanding the roles of proteins, in particular how they interact with each other, is the focus of this thesis, and a key step to understanding the whole operation of the cell.

2.2.1 From DNA to Proteins

Cells are the fundamental working units of every living system. The nucleus of every cell in eukaryotic organisms (including animals and plants) contains a large DNA (Deoxyribonucleic acid) molecule, which carries the genetic information of every organism.

DNA consists of two long chains of nucleotides. Each nucleotide is composed of one sugar molecule, one phosphate molecule, and a nitrogenous base. Four different bases are present in DNA: adenine (A), thymine (T), cytosine (C), and guanine (G). The particular order of the bases in any of the DNA strands is called the DNA sequence. The two DNA strands are complementary, which means that they contain the same genetic information (the information is duplicated) and are held together by weak hydrogen bonds.

The DNA sequence contains instructions for the synthesis of every protein. These are the specific sections of the DNA sequence usually called *genes*. The way how the information stored in the DNA passed on for the synthesis of proteins is called the central dogma of molecular biology. A simplified scheme of this process can be seen in Figure 2.1. This is commonly represented by two

main steps as follows:

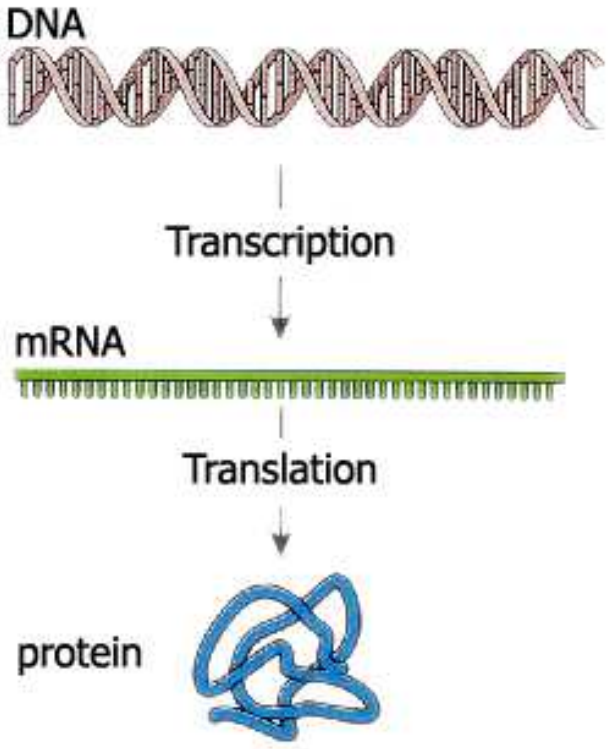


Figure 2.1: Scheme of the central dogma of molecular biology. (Figure adapted from *http://encephalon.ca/?p=4*)

i Transcription (DNA → mRNA): is the process where the information coded in a specific segment of the DNA sequence (or gene) is passed to a RNA molecule called messenger RNA (mRNA). RNA molecules are similar to DNA. They are also a chain of nucleotides, but contain only one strand and use different nitrogenous bases and sugars. Additionally, mRNA is smaller due contains the information related to only one gene. The process by which genes are transcribed into a RNA molecule is usually called gene expression.

ii Translation (RNA → Protein): is the process where the genetic information now coded in the mRNA is used to synthesize a specific protein.

This process is mediated by other macromolecules called ribosomes and also other types of RNA molecules. The genetic information is translated from a chain of nucleotides from the (mRNA) to a chain of amino acids. This is made using the genetic code, where a nucleotide triplet (codon) is associated with a specific amino acid. There are a total of 20 different amino acids. The final sequence of amino acids generated corresponds to what we know as a protein.

2.2.2 Protein Structure

Proteins are polymers consisting of chains of amino acids. The structure and shape of the proteins (how the chain of amino acids folds in 3-dimensional space) is relevant to determine their specific function.

Protein structure can be described at various levels. The first level is called the primary structure and corresponds to the linear amino acid sequence. The secondary structure refers to how the amino acid back bone of the protein is arranged in 3-dimensional space, by forming hydrogen bonds with itself. There are three main components in the secondary structure: alpha helices, beta sheets and random coils. The tertiary structure is produced when elements of the secondary structure fold up among them. Finally, the quaternary structure is related to the spatial arrangement of several proteins. Figure 2.2 present a schematic representation of these structural conformations. The final protein structure determines the function of each protein. More details about this can be found in (Shoemaker and Panchenko 2007a). In this thesis, we intend to employ information related to all these protein structure levels to infer PPI.

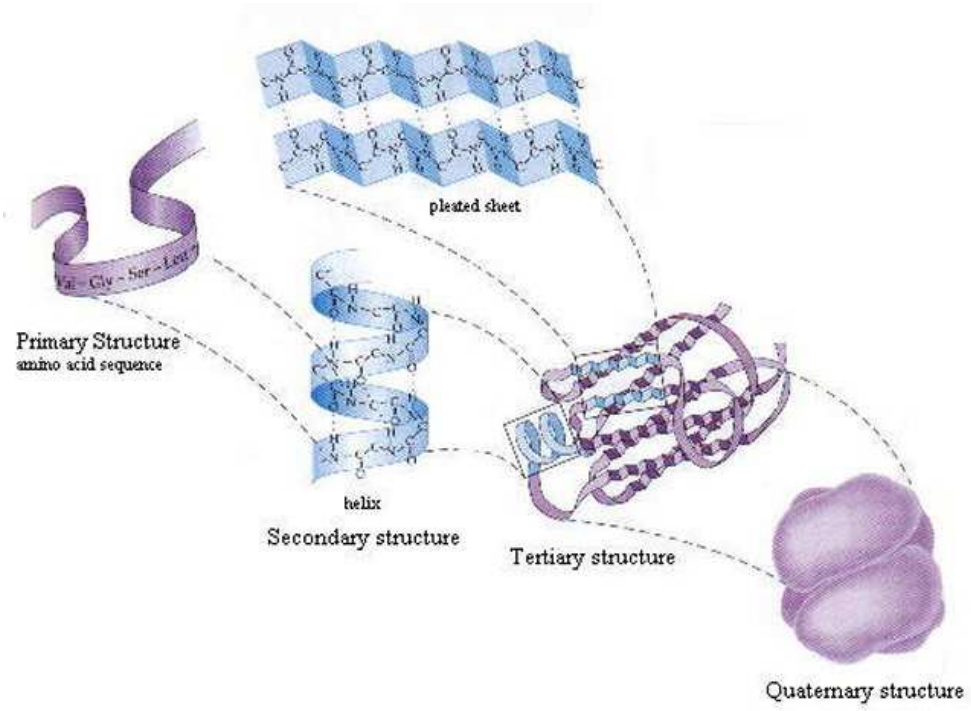


Figure 2.2: Scheme of different protein structure levels. (Figure adapted from *http://stevebambas.com/AP 220 Chemistry.htm*)

2.2.3 Protein Function

Proteins are involved in almost all the functions performed in a cell. Among these we find:

- enzymes which catalyze many metabolic reactions
- structural proteins such as those present for instance in the cell wall
- regulatory proteins, such as transcription factors that regulates the transcription of genes
- signalling molecules such as certain hormones like insulin.

Nowadays, due to the availability of high-throughput sequencing techniques, we know the complete genome sequence (DNA) of several species. Through this, we are also able to obtain the amino acid sequence of most proteins. However, the

function of a large portion of these proteins remains unknown, and consequently the inference of protein functions is still one of the most important research areas in bioinformatics.

The study of protein-protein interactions can potentially help with this task. If we are able to predict new PPI, then we could infer, for instance, the unknown function of certain proteins which interact with known ones (Shoemaker and Panchenko 2007a). Thus, the study of PPI could help us to understand how protein functions within the cell.

2.3 Protein-Protein Interactions

As previously mentioned, proteins usually do not work alone but in coordination with other proteins. This generates binary protein-protein interactions (PPI), protein complexes and protein interaction networks. Thus, these interactions control, regulate and participate in most cellular processes. These processes include cell cycle control, cell differentiation, protein folding, signaling, transcription, translation, post-translational modification and transport.

Considering the task of prediction of PPI, we can distinguish three main general kinds of PPI which have been studied in previous investigations:

- Protein complexes: these are related to proteins which are members of the same protein complex. In this case, any two proteins interact with each other if they are members of the same complex. In general, these interactions are more related to stable interactions.

- Physical interactions: these are related to direct interactions between two proteins which can occur at any time.

- Protein interaction networks: where proteins involved in binary and co-complexed interactions within an organism can be grouped and viewed as

a whole system.

In this thesis we will focus mainly on the predictive task associated with the first case related to protein complexes. Additionally, we will address a preliminary analysis of the PPI networks discovered.

Protein interactions can also be classified based on a number of different characteristics (Shoemaker and Panchenko 2007a):

- Strength: PPI can be stable or transient according to its life expectancy. Stable interactions are more related to groups of proteins that work together forming complexes. Transient PPI seems to be associated to the control of cellular processes.

- Specificity: interactions can be specific or non-specific. A specific interaction means that one protein can only interact with another specific protein partner.

- Similarity between interacting subunits: PPI are classified as homo-oligomers or hetero-oligomers, depending on whether the protein subunits involved are of the same type or not.

The prediction of PPI types based on some of these classification characteristics will be addressed in the last part of this thesis.

2.3.1 Identification/Detection of PPI

Protein-protein interactions are of central importance for most process in living organisms. Thus, information about these interactions can help to improve our understanding of diseases and can also serve as the basis for new therapeutic approaches (Shoemaker and Panchenko 2007a). Several experimental techniques have been developed in the past for the direct detection of PPI. Additionally

indirect detection approaches has been studied in the past, based on different types of biological information.

Small Scale Experiments

PPI interactions can be studied individually by using small scale experiments. These are based on the use of of genetic, biochemical and biophysical techniques (Phizicky and Fields 1995). These experiments are performed by measuring the natural affinity of binding partners utilizing either *in vitro* or *in vivo* approaches.

In vitro methods: This type of technique is developed in a controlled environment outside a living organism. A list of the most common *in vitro* methods is given in Table 2.1. All these methods exhibit advantages and disadvantages and generally provide complementary information.

Table 2.1: Commonly employed in vitro experimental methods for detection of PPI

Method	Description
Protein Arrays	Antibody-based or bait-based arrays detect interactions of proteins from complexes mixtures
Co-Immunoprecipitation	A purification procedure to determine if two different proteins interact
FRET	Fluorescence Resonance Energy Transfer (FRET) studies the transfer of two interacting proteins carrying fluorescence labels
NMR	Nuclear Magnetic Resonance (NMR) provides insights into the interaction of proteins in solution
X-ray Crystallography	Crystallization of an interacting complex. Allows definition of the interaction structure

In vivo methods: In this case, the experimental technique is developed inside the organism. The most widely used *in vivo* method to detect PPI is the "yeast two hybrid" (Y2H) system. The Y2H utilizes the transcription process

to identify protein interactions. Interactions detected by this approach often require confirmation from *in vitro* techniques to improve confidence.

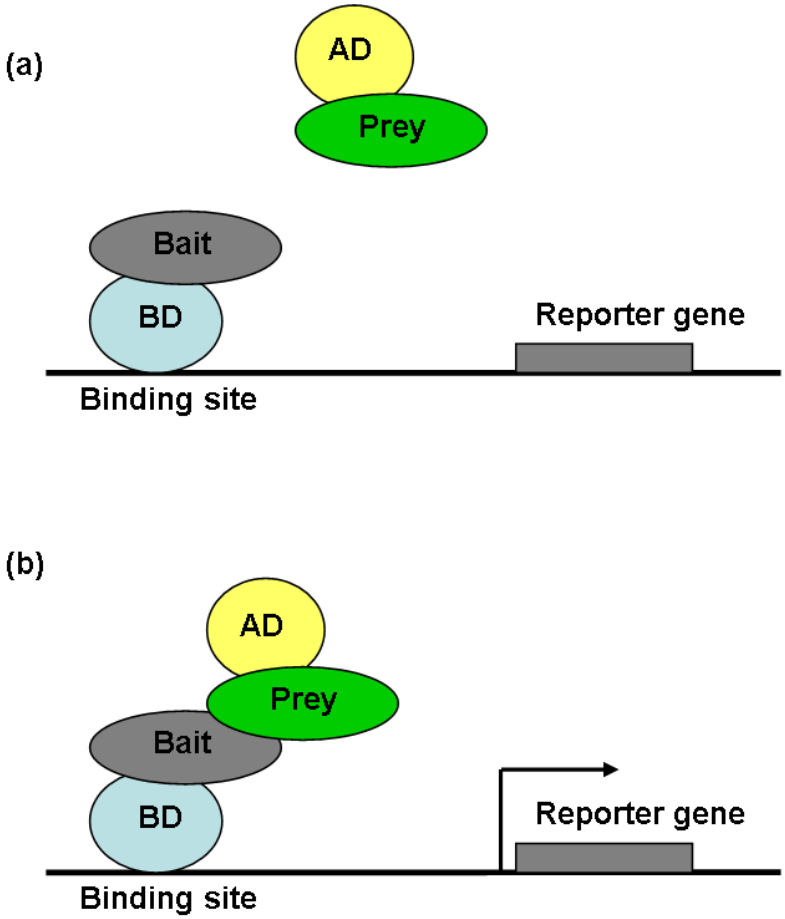


Figure 2.3: The yeast-two-hybrid system. (a) The DNA-binding domain hybrid does not activate transcription if protein "Bait" does not contain an activation domain. The activation domain hybrid does not activate transcription either because it does not localize the DNA-binding site. (b) Interaction between "Bait" and "Prey" brings the activation domain into close proximity to the DNA-binding site and results in transcription of a reporter gene.

The principle of the Y2H method is described in Figure 2.3. Pair of proteins to be tested for interaction are expressed as fusion proteins (hybrids) in yeast. The bait protein is fused to a transcription factor DNA binding domain. The other protein, the prey protein, is fused to a transcription factor activation domain. When expressed in a yeast cell containing the appropriate reporter

gene, interaction of the bait with the pray brings the DNA binding domain and the activation domain in to close proximity, creating a functional transcription factor. This triggers transcription of the reporter gene. The interaction can then be detected by expression of the linked reporter genes (Phizicky and Fields 1995, Shoemaker and Panchenko 2007a).

The Y2H technique has been used extensively both on the large-scale and for individual interaction experiments. It has been successfully applied to several organisms.

Small scale experiments are very expensive and time consuming. Consequently, most of the protein-protein interactions have not been discovered and validated experimentally. It is not possible to study all possible interactions between two or more proteins. However, at the moment this is the most accurate and reliable option for the correct detection of PPI. Thus, the main challenge seems to be in the selection of potential targets (two proteins with more chances to interact) to be studied employing small scale experiments. In order to address this challenge, many computational approaches have been proposed in the past, based on the use of machine learning techniques for the prediction of new PPI. These approaches will be reviewed in the next chapter.

Large Scale Experiments

The speed at which new proteins are discovered or predicted has created a need for methods that can detect high-throughput or large scale interaction data. In recent years, methods that can tackle this problem have been developed and introduced, resulting in a vast amount of new interaction data (von Mering et al. 2002). The two most popular types of large scale approaches are as follows:

Yeast two hybrid (Y2H) assay: high-throughput Y2H screens are based on the same principle discussed in the previous section (small scale / *in vivo* methods), but applied to entire genomes at the same time. This is used to detect pairwise binary interactions systematically on a large scale.

The first two Y2H analyses were carried out in yeast and revealed 841 and 692 putative interactions respectively (Uetz et al. 2000, Ito et al. 2001). The overlap between these two investigations results was very small. Only 141 interactions (around 20%). Y2H screens have been recently employed for other organisms such as fly, worm and human (Rual et al. 2005). This approach is specifically useful for the prediction of transient and unstable interactions. However, this technique could easily miss certain interactions due to insufficient depth of screening and misfolding of the fusion proteins. In addition, the process associated takes place in the nucleus, so many proteins are not in their native compartment.

Mass spectrometry methods: Protein complex purification and identification techniques using mass spectrometry (Gavin et al. 2002, Ho et al. 2002, Gavin et al. 2006) are employed to reveal the components of protein complexes, i.e. multiple proteins that interact with each other mostly directly but also indirectly. The general process of this type of method has four main steps: (a) Individual proteins are tagged and employed as baits to biochemically purify whole protein complexes. (b) Bait proteins are systematically precipitated, along with any associated proteins, on an "affinity column". (c) Purified protein complexes are resolved by one-dimensional SDS-PAGE, a technique that involves running an electric charge through the complexes on a gel, so that proteins become separated according to mass. (d) Proteins are excised from the gel, digested with an enzyme, typically trypsin, and the digest is analyzed by mass spectrometry.

Data base-search algorithms are finally used to identify specific proteins from their mass spectra.

For large-scale mass spectrometry based protein complex purification techniques, their advantages include: several members of a complex can be tagged at once by this technique, and it detects real complexes in physiological settings. However, these methods may miss complexes that are not present under the given conditions. Also tagging may disturb complex formation, and weakly associated components may dissociate and scape detection.

In general, the interaction data generated by large scale techniques is incomplete and noisy, as well as being difficult to reproduce. For these reasons such studies are frequently criticized in the literature. In this case, the challenge seems to be how to improve the accuracy and reliability of the PPI inferred by large scale experiments.

2.4 Availability of PPI Data

In recent years much work has been carried out in order to improve our knowledge of PPI. However only a small fraction of the total of PPI has been trustworthily identified. Currently available PPI data generated using experimental techniques is still preliminary in terms of quality as well as quantity.

The work in von Mering et al. (von Mering et al. 2002) was the first to undertake a comprehensive analysis to compare different sets of PPI inferred for yeast. The data analyzed corresponds to PPI detected by using small scale as well as large scale experiments. They measure the accuracy and identify biases, strengths and weaknesses for the different approaches. Their results showed that from among approximately 80,000 interactions between yeast proteins from different high-throughput methods, only a small portion (around 2,400 PPI) were

supported by more than one method. This suggested that either the methods identify different sets of PPI, or the methods produced a high portion of false positives. In fact they estimated that false positive and false negative rates associated with high-throughput techniques is around 50%.

In this thesis, we focus in the study of PPI associated with yeast. Several databases have been created in recent years to compile information about PPI (Shoemaker and Panchenko 2007b), including data from diverse experiments reported in the literature as well as manually annotated PPI data sets. Table 2.2 exhibits a list of PPI databases which are of public access for researchers. Among them we found the Munich Information Center for Protein Sequences (MIPS) (Mewes et al. 2002) database, The IntAct database (Kerrien et al. 2007), The Molecular Interactions (MINTS) (Shoemaker and Panchenko 2007a) database, the Database of Integrating Proteins (DIP), the Biomolecular Interaction network Database (BIND) (Bader et al. 2001), and the BioGRID (Reguly et al. 2006) database.

Table 2.2: List of public PPI databases

Database name	Number of PPI	URL
MIPS	15,488	http://www.mips.gsf.de/services/ppi
IntAct	68,165	http://www.ebi.ac.uk/intact
BioGRID	116,000	http://www.thebiogrid.org
DIP	55,733	http://www.dip.doe-mlb.ucla.edu
BIND	83,517	http:/www.bind.ca
MINT	71,854	http://www.mint.bio.uniroma2.it/mint
STRING	730,000	http://string.embl.de

2.5 Summary

In this chapter we have presented an overview of protein-protein interactions from a biological perspective. We first described the basic biological topics associated, from DNA to protein interaction networks. We then described the main experimental tools available today for the identification of PPI. Finally the major PPI databases were reviewed.

Existent experimental technologies for the identification of PPI exhibit many limitations. Consequently, available interaction data sets are incomplete and highly noisy. In order to deal with this situation several machine learning techniques have been recently employed to deal with the problem of PPI prediction. In the next chapter we will provide an overview of the work associated with this prediction task.

Chapter 3

Related Work

3.1 Introduction

In this chapter we will present an overview of machine learning techniques employed in the past to deal with the problem of prediction of PPI. Firstly, we will present a brief overview of the machine learning field, including the description of several learning algorithms that will be employed in this thesis. Then we will introduce the concept of "learning from diverse types of data", describing a general framework related to how machine learning algorithms can be implemented and applied over heterogeneous sources of data, in order to develop a joint integrative analysis. Furthermore we will present a detailed review of machine learning approaches utilized in recent years to deal with the problem of PPI prediction. Finally we will introduce fundamental issues associated to the problem of prediction of PPI.

3.2 Machine Learning Overview

3.2.1 Learning Process

Machine learning (ML) can be defined as the study of computational methods and the construction of computer algorithms and programs capable of learning from their own previous experience, in order to improve their performance at a defined task (Mitchell 1997).

This field is usually related to other research areas, such as pattern recognition and statistical inference (Mjolsness and Decoste 2001), and also must be considered as a multi-disciplinary field that applies ideas from different areas such as: artificial intelligence (AI), probability, statistics, information theory and signal processing, computational mathematics, philosophy, control systems theory, cognitive psychology, biology, economics, operations research (OR) and others.

The concept of learning, according to Mitchell, 1997 (Mitchell 1997), is related to "acquiring the definition of a general category given a sample of positive and negative training examples of the category". This can be used, for example, in the context of finding the hypothesis that best fits the training examples in a defined space of hypotheses. More simply, the concept of learning can be related to the incorporation of new information or knowledge from the training examples into the system being studied.

In the definition of a learning problem, there are three main components to consider, which are: (1) The Training Experience (E) related to the training data sets from the system will learn; (2) The Class of Tasks (T), related to the definition of the target function that determines the type of knowledge will be learned; (3) The Performance Measure (P) of the knowledge that is acquired in the process.

It is possible to recognize two main categories of learning:

- *Supervised learning:* in this case the goal is to predict the value of an outcome measure based on a number of input measures, in which both the inputs and the outputs can be observed (Hastie et al. 2003). The principal tasks associated with this kind of learning are: Classification, Regression and Prediction.

- *Unsupervised learning:* in this case there is no outcome measure and the goal is to describe the associations and patterns among a set of input measures (Hastie et al. 2003). The principal tasks associated with this kind of learning are: Clustering and Association Rules.

In this thesis we focus in the supervised classification task. Classification attempts to divide the data into classes. A characterization of the classes can then be used to make predictions for new unclassified data. Classes can be a simple binary partition (such as a pair of proteins "interact" or "not interact" for the problem we face in this thesis), or can be complex with multiple classes as in the prediction of gene functional hierarchies.

To perform any machine learning task, there are general steps one must perform for successful pattern recognition. This mainly involves collecting data (variables and features), performing feature selection (for instance removing irrelevant and redundant features), choosing the right learning algorithm for your data (for instance evaluating several alternatives), training the classifier (or model), and finally evaluating the performance of the classifier (usually performed on a separate test set). A typical machine learning approach for classification is given in figure 3.1.

Performing feature selection is a critical step in the classification process. With a large data set and high-dimensional feature vectors, it would be expected

that the classifier would perform poorly due to the redundant and irrelevant features present in the training set. However, by selecting features that are invariant to irrelevant transformations, insensitive to noise and highly discriminatory then we could expect to achieve a more successful pattern recognition model.

The choice of a learning algorithm is also an important step. For instance, some methods such as Support Vector machines (SVM) (Vapnik 1998) are very flexible to deal with high dimensionality. Some learning algorithms are severely affected by the imbalanced data problem such as SVM and Decisions Trees(DT), while others like Naive Bayes (NB) are not. Some machine learning algorithms produce human readable results whereas others are "black boxes", whose working and intuition can not be understood. SVM is and example of a black box approach. However they are often highly accurate in their results, particularly on continuous real-valued numeric data. In this thesis we will evaluate diverse types of learning algorithms.

After training a classifier, the classifier performance is measured by applying an evaluation procedure. Many statistical and other measurements exits in Machine Learning. One obstacle that might affect the evaluation procedure is overfitting. This occurs when a classifier allows for perfect classification on the training data while performing poorly on a new data set (test data). A common way of overcoming this situation is to provide an independent test data set (validation set). While training on the training examples, the learning algorithm will monitor the error on the training set with respect to the validation set, and thus adjust the performance of the classifier accordingly. However This is possible mostly when a large amount of data is available. On smaller amounts of data, holding out a large enough independent test set may imply that not enough data is available for training. On these case a common solutions is to perform a cross validation procedure which will be explained in detail later in this chapter.

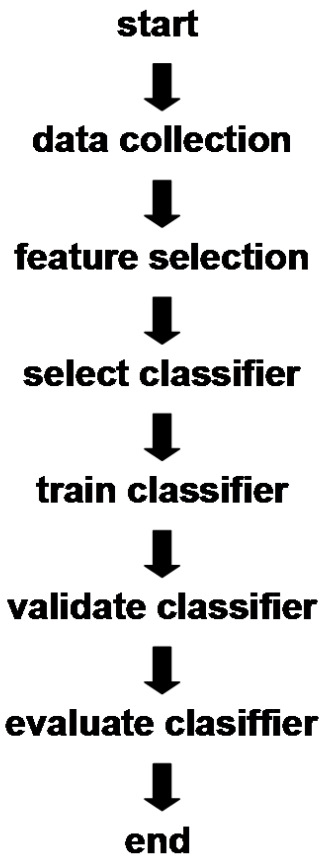


Figure 3.1: A typical machine learning process for classification

3.2.2 Conventional Machine Learning Algorithms

Here we describe several conventional machine learning techniques that have been used in the past for the task of PPI prediction. These approaches were also employed in this thesis to develop a comparative performance evaluation analysis.

Decision Trees: The decision tree is a supervised learning technique that uses approximating discrete functions to estimate and classify the examples. In the nodes of trees are attributes and in the leaves are values of discrete function. The decision tree can be rewritten in a set of "*if-then*" rules and also give an estimation of the probability of occurrence of a particular case. This is an

inductive learning method which is very popular and mostly used for variety of classification tasks.

Naive Bayes: is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In spite of their naive design and apparently over-simplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect.

Support Vector Machines (SVM): The Support Vector Machines were developed by Vapnik and co-workers (Vapnik 1998), based on the Structural Risk Minimization principle from statistical learning theory. This is a supervised learning method, mainly applied to classification and regression problems. The main idea of an SVM is to separate classes with a surface that maximizes the margins between them. This method combines two main ideas. The first is the concept of an optimum linear margin classifier that constructs a separating hyperplane that maximizes distances to the training point. This hyperplane is supported by some of these training points. The second is the concept of a kernel which is a function that calculates the dot product of two training vectors. Kernels calculate these dot products in feature space. When using feature transformation, which reformulates input vectors into new features, the dot product is calculated in feature space, even if the new feature space has higher dimen-

sionality. The linear classifier is unaffected.

Logistic regression: is a model used for prediction of the probability of occurrence of an event by fitting data to a logistic curve. It makes use of several predictor variables that may be either numerical or categorical. For example, the probability that a person has a heart attack within a specified time period might be predicted from knowledge of the person's age, sex and body mass index. Logistic regression is used extensively in the medical and social sciences as well as marketing applications such as prediction of a customer's propensity to purchase a product or cease a subscription.

3.3 Learning from Diverse Types of Data

In recent years there has been a rapid growth in the generation and storage of large and diverse data sets related to many scientific and commercial disciplines. Among them we can mention: business information, marketing and sales data, medical records, biology and other scientific databases (Caragea et al. 2004, Friedman et al. 1999, Getoor et al. 2001, Dzeroski 2003). This has been possible mainly due to the availability of new high-throughput data acquisition methods and advances in computing, communications and digital storage technologies. For example, organizations have begun to capture and store a variety of data about various aspects of their operations (e.g. products, customers, and transactions). On the other hand, most of the productive processes are coupled with complex distributed systems such as computer systems, communication networks and power systems, which are equipped with sensors and measurement devices that collect and store a variety of data, for use in monitoring controlling and improving the operation of such processes.

These large data sets are usually stored in different, autonomously struc-

tured and relational data repositories. This means that objects in these diverse data sets often have a rich internal structure and are connected by some relation. The common reason why this information is stored independently is usually for storage space and retrieval efficiency considerations (i.e. distributed databases) (Merugu and Ghosh 2005), which results better when dealing with relatively small data sets linked in some specific manner. In other cases, the generation of diverse and heterogeneous data sets related to a specific area depend on other factors such as: the different data sets representing different insights about the same field (i.e. biological databases include DNA sequences, protein sequences, gene expression data, etc); each data set being generated by different experimental methods so they do not result in the same accuracy; different data sets presenting a heterogeneous distribution and representation. Due these reasons it is not possible to store them in single table (Ben-David et al. 2002).

The availability of huge amounts of data represents an important opportunity for large-scale data-drive knowledge acquisition (Caragea et al. 2004), especially in scientific areas where it seems to be possible gain a deeper understanding in many data-rich domains. Consequently, the integrative analysis of diverse and relational data sets has become an emergent area of research, with special emphasis on mining these databases in order to look for patterns and associations that allow us to improve our understanding in these areas and discover useful relationships (Friedman et al. 1999, Getoor et al. 2001, Merugu and Ghosh 2005, Ben-David et al. 2002). This is especially important in fields like bioinformatics and computational biology, where the relations between different kinds of information (represented in different datasets) are not previously known and so the integrative analysis of these diverse data sets could potentially reveal novel aspects of biological systems (Kanehisa and Bork 2003, Filkov and Skiena 2004).

Unfortunately, most of the common existing machine learning and data mining approaches are restricted to dealing with data stored in a single relation of a database (Friedman et al. 1999, Getoor et al. 2002, Dzeroski 2003, Domingos 2003), where the instances are independent and identically distributed (i.i.d.), and so they are not able to deal with multiple heterogeneous data sources in a direct way. In this sense, it is necessary to develop new machine learning techniques or an extension of the common ones in order to deal with multiple sources of data.

In order to solve this problem, a traditional and simple way of mining multiple databases is typically to integrate the diverse data sets into a single table, and then apply some machine learning and data mining techniques to this new joint data set in order to generate a new knowledge process. This process is generally known as *data integration* (Dhamankar et al. 2004, Rahm and Bernstein 2001), and has been mainly applied to schema matching, which is the problem of producing semantic mappings which transform data instances from one schema to instances of another.

In relation to the integration of biological databases, recent reviews about current systems and challenges in this area can be found in (Wong 2002, Stein 2003, Hernandez and Kambhampati 2004).

The integration of diverse and heterogeneous data sets is often hard and in general presents several important disadvantages most of them related to the potential performance of the machine learning integrative analysis (Caragea et al. 2004, Yin et al. 2004, Reinoso et al. 2003), as follow:

- due to the large size of the diverse data sets, gathering all the data in a centralized location in general is not desirable and some times not feasible owing to storage and privacy requirements

- the learning process time increases significantly with the data size

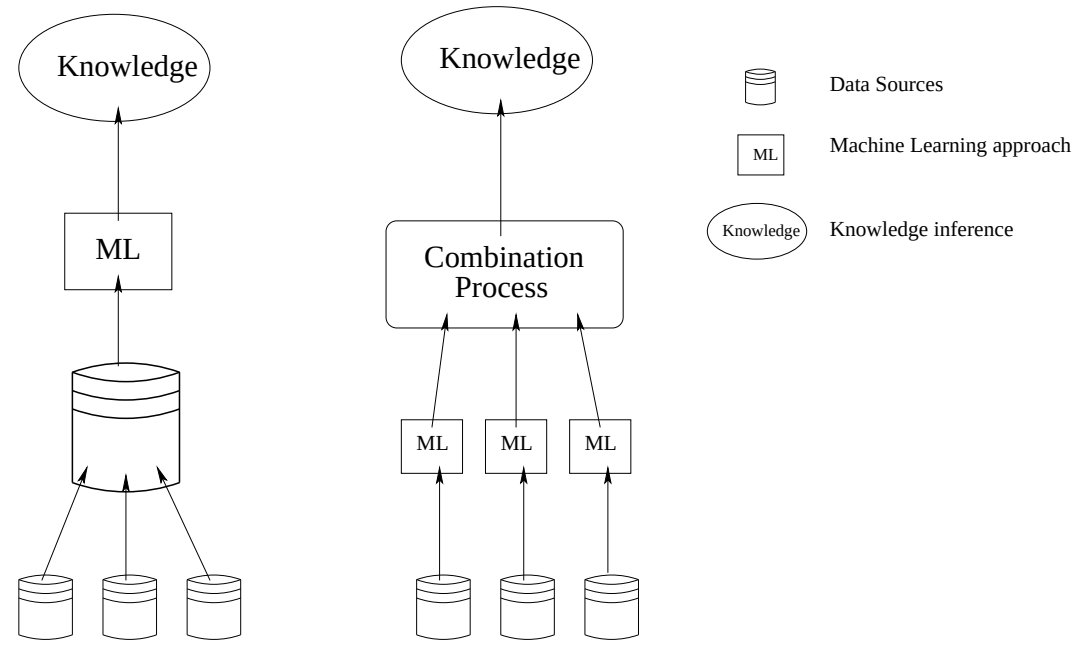
- the integration process can introduce duplication and redundancy
- the integration process removes the structure of the diverse data sets, losing information, which could be crucial for the objective of discovering hidden knowledge.

Consequently, the necessity to develop or implement novel machine learning methods for analyzing and mining heterogeneous and diverse data sources has started to receive more attention in recent years.

3.3.1 General Integrative Framework

It is important to identify a general framework related to how machine learning algorithms can be implemented and applied over heterogeneous sources of data, in order to develop a joint integrative analysis. In this sense, it is possible to classify two main general approaches (see figure 3.2 for an schematic representation of both approaches):

- the simpler approach is the direct integration of different databases in order to generate a new unified data set. Then it is possible to apply some machine learning techniques in order to learn and discover new knowledge from this unified data set. This approach corresponds to the method previously presented as *data integration*, which highlights various disadvantages related to the potential performance of the machine learning integrative analysis. The main problem is the possible loss of information as a product of the elimination of the structure in the integration process. Consequently, it is not possible to infer all the relations between the different data sets and therefore results in the generation of an incomplete model that does not have all the potential knowledge (Getoor et al. 2001).



(a) Data Integration approach (b) Ensemble related approach

Figure 3.2: Scheme of general approaches for the application of machine learning algorithms over heterogeneous sources of data

This approach must be considered as a baseline in order to compare with other possible approaches. For example, when any other machine learning based approach is developed to analyse heterogeneous data sources, the performance of this learning method (e.g. accuracy and/or interpretability) must be at least better than the application of the same machine learning concept to the simple integration of different data sources.

- a second approach is to treat each data source separately, and to use some machine learning techniques to generate independent models that represent and allow us to obtain separate inferences and knowledge from each data type. After that, the idea is to generate a unified model (a combination of the previous ones) that represents the information and knowledge of the whole data set. This unified model is the result of the combination of the diverse independent learning models at different levels:

i Combination of knowledge, where information is captured from the diverse data sources and then combined in order to perform general inferences, (i.e. instances are classified by each of the independent learning models and then the final classification is the combination, in some way, of these independent classifications).

ii Combination of learning models where the idea is to combine the different learning models (i.e. classification models) in some way, in order to generate a unified model and then use this unified model over the instances to perform general inferences (i.e. instances are classified using the unified model).

In both cases, the idea of combination is similar to the *ensemble machine learning framework*, where the idea is to learn separate classifiers individually induced from diverse data sources and combine them in some way at the end of the learning process. Many studies of ensemble methods have been developed and are summarized in (Ali and Pazzani 1996, Bauer and Kohavi 1999, Dietterich 2000b).

Both approaches will be employed and evaluated in this thesis as will be explained in detail in Chapters 4 and 5 respectively.

3.3.2 Applications in Bioinformatics

Recent developments in high-throughput techniques have generated a wide variety of different sources of biological information at a genome-wide scale, including many kinds of data, such as: DNA sequences, gene expression data, protein sequences, protein-protein interactions, protein-DNA binding data, protein structural information, phylogenetic profiles, metabolic data, physiology data, and several others.

Machine learning techniques have been broadly applied for research in the field of bioinformatics and/or computational biology (Baldi and Brunak 2001, Narayanan et al. 2002, Keedwell and Narayanan 2005). However, most of this work has been dedicated to the analysis of a single type of data at a time, using other types of data only for validation. In contrast, results of the joint analysis approach from more than one type of data allow for the finding of new insights that may not be as readily available from analyzing one type of data in isolation.

The integrative analysis of these diverse biological data is an emerging and important issue in bioinformatics and computational biology research, considering that each one of these distinct types of data provide a particular view of the molecular machinery of the cell and probably contain different and thus partly independent information. By combining those complementary pieces of information, it could be possible to enhance our knowledge about the relationships between the different components of a genome, to discover new biological insights that may not be as readily available from analyzing one type of data in isolation and therefore improve the results of the previous analysis (Lanckriet et al. 2004). A major challenge in this sense is to develop a unified framework for combining the multiple sources of biological data of an organism, and to look for associations between them, thus obtaining a robust and integrated view of the underlying biology, which can be also considered as a Systems Biology approach (Wolkenhauer and Gierl 2003).

The main advantages of using an integrative approach in the field of bioinformatics can be summarized in the following list:

- dealing with errors contained in experimental data. Biological datasets often contain errors arising from imperfections in the applied technology. If we assume that technological errors across different datasets are largely independent, then the probability of error in results that are supported

by two or more different types of data is significantly reduced (Hartemink and Segal 2005).

- improving the results of previous analysis made using just one kind of data. For example, in the case of prediction of protein function, the use of an integrative approach increases the prediction accuracy from 57% to 87% (Deng et al. 2003).

- The integrative analysis of all kinds of biological data available allows to us to discover new knowledge and thus obtain a better understanding of complex biological systems and processes. This is not possible with the analysis of just one kind of data (Wolkenhauer and Gierl 2003).

3.4 Machine Learning for Prediction of PPI

Protein-protein interactions play a key role in most biological processes (Szilágyi et al. 2005). Thus its identification can help to understand the working mechanisms of the cell. Previously it has been pointed out that high-throughput experimental interaction data exhibit high false positive and false-negative rates (von Mering et al. 2002, Edwards et al. 2002). On the other hand small-scale experiments are costly and time consuming. Consequently, most of real PPI have not been discovered experimentally.

In addition to experimental information based on direct interaction evidence. In recent years many investigations have been focused in the analysis of indirect sources of evidence about PPIs, with the aim to improve the inference of PPI pairs. These include:

- it has been demonstrated that in general protein interacting pairs exhibit similar expression profiles (are co-expressed among different conditions) (von Mering et al. 2002).

- The expression of protein members of the same complex frequently are regulated by the same transcription factors (Bar-Joseph et al. 21).

- Protein sequence data has been employed to infer PPI, based in the over-representation of certain domain or motifs pairs (Gomez et al. 2003, Deng et al. 2002, Wang et al. 2005, Wang et al. 2007, Li et al. 2007, Wu and Zhang 2008).

- Protein structure information have also been incorporated for this task. Although mostly imitated by the number of available PPI structures (Espadaler et al. 2005, Chia and Kolatkar 2004).

Based on the above observations, a number of researchers have suggested that direct data on protein interactions can be with indirect data in a supervised learning framework (Bader et al. 2004, Gilchrist et al. 2004, Jansen et al. 2003, Lin et al. 2004, Zhang et al. 2004, Ben-Hur and Noble 2005, von Mering et al. 2005) Investigations employing this general approach generally use a certain classification algorithm to integrate diverse biological data sets. The classifier is trained to distinguish between positive (truly interacting protein pairs) from a set of negative examples of non-interacting pairs. It is important to mention that most of these approaches have utilized the "data integration" general approach described previously in figure 3.2.

The problem of accurately inferring PPI from high-throughput data was firstly discussed in Von Mering et al. (von Mering et al. 2002). The solution proposed there is based in the intersection of several high-throughput experimental data, achieving a low false-positive rate, but a very low coverage at the same time. The "STRING" database built by the same authors (von Mering et al. 2005) integrates protein interaction evidence derived from high-throughput experimental data, from the mining of databases and literature, and from pre-

dictions based on a genomic context analysis. The work in (Jansen et al. 2003) proposed the use of a naive Bayes classifier based on examples derived from the MIPS database. The approach in (Lin et al. 2004) extended the previous work evaluating other classifiers. They also discussed the importance of diverse features and concluded that protein functional information was the most informative. Logistic Regression (LR) was employed in (Bader et al. 2004) to estimate the posterior probability that a pair of proteins will interact, based on several high-throughput experimental data. An approach based on Decision Trees (DT) was employed in (Zhang et al. 2004), again based in the MIPS protein complexes database.

When analyzing PPI networks at the level of binary interactions, much information is lost, because protein often perform their functions together in groups. Understanding these interaction groups called complexes, is essential for systematically modeling the behavior of cellular networks.

Graph analysis algorithms can help us to understand how proteins are logically connected. The connection between proteins can be represented on an indirect graph, where the nodes correspond to proteins and the edges correspond to the interactions. Thus the identification of new complexes can be simply viewed as the computational problem of locating important subgraphs. This kind of analysis can produce valuable insights, considering the topological properties as well as the functional organizations of protein networks in cells.

Many cellular functions are performed by complexes containing multiple protein interaction partners. Predicting molecular complexes, one of the fundamental units in PPI networks, is one of the most important tasks in the analysis of protein interaction networks. High-throughput experimental approaches to identify yeast protein complexes on a proteome-wide scale has been employed in the past (Ho et al. 2002, Gavin et al. 2002, Gavin et al. 2006). As previ-

ously stated these kind of information suffers from high false positive and false negative rates (von Mering et al. 2002). In order to deal with this situation, there have been various computational attempts to accurately identify protein complexes from this type of data (King et al. 2004, Dunn et al. 2005, Pereira-Leal et al. 2004, Bader and Hogue 2003, Adamcsek et al. 2006, Spirin and Mirny 2003, Rives and Galitski 2003, Arnau et al. 2005, Sharan et al. 2005, Scholtens et al. 2005, Chu et al. 2006). These methods have mostly employed an unsupervised graph clustering approach, aiming to discover similarly or densely connected subgraphs (clusters) (Aittokallio and Schwikowski 2006).

In general the methods previously mentioned have presumed that protein complexes correspond to the dense regions of networks. However this not true in all cases, there are other topological structures that may also represent a complex. One example is the called "hub" or "star" model, in which many vertices (proteins) connect to a central "hub" protein (Bader et al. 2004). Another interesting topology is a structure that links several small connected components, which can be associated to large protein complexes (Qi et al. 2008).

3.4.1 Machine learning Issues Important for the Prediction of PPI

The problem of prediction of PPI has been classically tackled as a binary classification problem in a supervised learning context. Where the objective is to generate a classification model able to predict whether any two proteins do or do not interact. The final goal associated to this computational approach is to use this predictive model to to predict or identify new potential PPI. These targets have to be investigated and validated later by biologists through the development of small scale experimental techniques for detection of PPI.

In this context we need a gold standard reference set of positive (truly in-

teracting protein pairs) and negative examples (non-interacting pairs). Positive and negative examples for this task are pair of proteins which can be represented by a *n-dimensional* vector X_i containing the information for the biological features considered here, and a label Y_i taking two values ($Y_i = 1$) when the pair of proteins really interact or ($Y_i = -1$) when there is not interaction. Each object or example is thus represented as one point in the *n-dimensional* feature space. We finally expect that biological information encoded in the feature vectors, can be helpful to characterize and discriminate positive from negative PPI examples.

In this research we will consider several assumptions as follows:

- *Complete information*: we assume that all available examples (positive and negative ones) in the gold standard reference set have complete information. There are no missing values and we assume all examples are characterized with the same set of features. In practice might happen that some measurements are not available. The missing values introduce extra complications and we will not consider this situation.

- *Continuity assumption*: Further, we assume that the continuity assumption holds. This is a general assumption in pattern recognition, two objects near in feature space should also resemble each other in real life. When we are at one position in feature space representing an example object, and we change the position a bit, then this new position should represent a very similar object. This also means that available examples in our task are not randomly scattered into some feature space, but they are distributed in a cloud-like distribution. When we look in the neighborhood of an object similar objects are represented.

In general a learning model can be represented by a function $f(X)$ (X in the *n-dimensional* feature space), which is able to assign an object X to one of the

classes $Y = 1$ (positive class) or $Y = -1$ (negative class). This is denoted in equation 3.1. This function is frequently chosen beforehand. Note that depending of the type of function selected, a set of parameters (W) associated to this function have to be determined (optimized) during the learning process.

For the case of prediction of PPI, this function divides the input space (*n-dimensional*) into two *decision regions*, one for each class. The boundary between these decision regions are called *decision boundaries or decision surfaces* (Bishop 2006). Function $f(X)$ is trained using the available examples denoted in equation 3.2. Such that a new example X is classified by this function into one of the available classes.

$$f(X) : \mathbb{R}^n \rightarrow \{\pm 1\} \tag{3.1}$$

$$(X_1, Y_1), ..., (X_N, Y_N) \in \mathbb{R}^n \times \{\pm 1\} \tag{3.2}$$

Where X_i are the N training examples represented as biological feature vectors in the *n-dimensional* space; Y_i are class labels (+1 or -1); and R^n is the *n-dimensional* feature space.

From a probabilistic perspective, the classification problem can be divided in two separate stages as shown in (Bishop 2006).

Firstly we consider the *inference stage*, where the training data is employed to model the joint probability distribution $p(X, Y_k)$. In the case of prediction of PPI The input vector X is a set of biological features and the output variable Y_k is the class label indicating whether two proteins do or do not interact (k takes values +1 or -1 in this binary classification problem).

This joint probability can be used later to estimate the conditional probabilities ($p(Y_k|X)$) of the two classes given a certain feature vector X. For instance

using the theorem of Bayes we can estimate these probabilities as follows:

$$p(Y_k|X) = \frac{p(X|Y_k)p(Y_k)}{p(X)} \tag{3.3}$$

It is possible to interpret $p(Y_k)$ as the prior probability for the class Y_k, and the $p(Y_k|X)$ as the corresponding posterior probability. Thus $p(Y_{k=+1})$ represents the probability that two proteins really interact without knowing information their biological information, and $p(Y_{k=+1}|X)$ is the corresponding posterior probability given the biological information.

The term $p(X|Y_k)$ is called the class likelihood and is evaluated directly from the observed training data set. When employing for instance the Naive Bayes learning approach, we consider that every attribute x_j in the feature vector X of dimension n are conditionally independent given the target value Y_k. Where $x_1, x_2, ..., x_n$ are the sequence of biological attributes (or features) in the vector X. Thus it is possible to estimate the term $p(X|Y_k)$ as shown in equation 3.4.

$$p(X|Y_k) = \prod_{j=1}^{n} p(x_j|Y_k)p(Y_k) \tag{3.4}$$

Secondly we consider the *decision stage*, where these posterior probabilities are employed to make optimal class assignments. For this, a decision function $f_{Bayes}(X)$ associated to this Bayesian approach is used to assign labels, where a new object (represented by a feature vector X) is assigned to the class with the largest posterior probability $p(Y_k|X)$ according the expression in equation 3.5. Note that the denominator in equation 3.3, $p(X)$, is common for every posterior probability, thus we are only interested in the numerator of these expressions when comparing them.

$$f_{Bayes}(X) = \begin{cases} +1 \ if \ p(Y_{k=+1}|X) \geq p(Y_{k=-1}|X) \\ -1 \ if \ p(Y_{k=+1}|X) < p(Y_{k=-1}|X) \end{cases} \tag{3.5}$$

An alternative solution to the classification problem is to solve both stages together and simply learn a function $f(X)$, called *discriminant function*, which maps each input X directly onto a class label Y_k, similar to the one in equation 3.1. This is the case of learning methods such as Support Vectors Machines (SVM).

An important objective of the model generated is to make as few misclassifications as possible. A mistake occurs when an object X belonging to the positive class is assigned to negative class or vice versa. A simple goal of a classification model could be to minimize the number of mistakes. This can be linked to an *error function* $\boldsymbol{Err}(f(X), Y)$ (also called *loss function*) which is a single, overall measure of loss incurred during the classification process (Bishop 2006). This loss function defines a measurable indicator of the miss-match between the model output $f(X)$ and the actual target value ($Y = [\pm 1]$) for all available objects (X_i, Y_i), where $i = 1, ..., N$.

Most often the objects in the training data (i.e. N examples in total) are assumed to be independently distributed, and the total error of function $f(X)$ on a training set is decomposed as in equation 3.6.

$$\boldsymbol{Err}(f(X), Y) = \frac{1}{N} \sum_{i=1}^{N} Err(f(X_i), Y_i) \tag{3.6}$$

There are different definitions for the error function, depending on the type of $f(X_i)$. The most simple is the called *zero-one loss* (Err_{0-1}) for discrete valued $f(X_i)$. basically this error function counts the number of wrongly classified objects (see equation 3.7

$$Err_{0-1}(f(X_i), Y_i) = \begin{cases} 0 \ , \ if \ f(X_i) = Y_i \\ 1 \ , \ otherwise \end{cases} \tag{3.7}$$

The most common error for real-valued functions $f(X_i) \in [\pm 1]$ is the mean squared error (MSE). The expression for MSE can be seen in equation 3.8.

$$Err_{MSE}(f(X_i), Y_i) = (f(X_i) - Y_i)^2 \tag{3.8}$$

By minimizing the error on the training set $\boldsymbol{Err}_{training}$, we hope to find a good classification model. However, this poses a new problem, the set of training examples might be a very uncharacteristic set. If a limited sample is available, the inherent variance in the objects and noise in the measurements might be too big to extract classification rules with high confidence.

In general, the larger the sample size, the better the characteristics of the data can be determined. But even when a good characteristic sample is available, there are many functions which approximates or precisely fits the data. Therefore, good classification of the training objects is not the main goal, but to obtain a good classification of new and unseen objects. How well a model trained on the training set predicts the right output for new instances is called *generalization* (Alpaydin 2004). The main goal in pattern recognition is to find classifiers that show good generalization (Bishop 2006).

To estimate how well a classification method generalizes, it has to be tested with a new set of objects, which has not been used for training. By using such an independent test set, we avoid an overly optimistic estimate of the performance. In many situations correctly labeled data is scarce and/or expensive. From these available objects both a training set as well as a testing set of objects should be drawn. Leaving out a set of objects from a reduced labeled set might leave out valuable information and therefore reduce the generalization of the classifier.

A way to overcoming this problem is to use a N-fold *cross-validation* procedure, where the training data is divide in N folds (N=10 is usually employed in machine learning). A portion (N-1)/N of the available data is used for training while the rest is employed to test the model. Finally this procedure is repeated N times to complete the process. This approach will be implemented in our research as will be seen in next chapters.

The phenomenon that a classifier allows for good classification on the training data (low $\boldsymbol{Err}_{training}$), while performing poorly on an independent test set (large $\boldsymbol{Err}_{test}$), is called overtraining or overfitting. This usually occurs when a too complex function or classification model $f(X)$ is employed. A sufficiently flexible function can always perfectly fit the training data and thus obtain a minimal $\mathbf{Err}_{training}$. The function then completely adapts to all available information, including noise in the given examples.

This overfitting problem becomes more sever when a large number of features is employed. Because the function $f(X)$ is defined for the complete feature space (i.e. *n-dimensional* feature vectors), the volume that should be described increases exponentially in the number of features n. This is called the *curse of dimensionality* (Bishop 1995). By decreasing the number of features per object, the number of degrees of freedom in the function f(X) decreased and the generalization performance increases. One solution to the curse of dimensionality and overfitting is to use feature reduction or feature selection and retain only the few best features.

As was stated before, the main goal in a classification process is to find classifiers that show good generalization. For this we have to be able to minimize the average of the error function $\boldsymbol{Err}(f(X), Y)$. In other words we want to minimize the *Expected error* ($\mathbb{E}\{\boldsymbol{Err}\}$) as is denoted in equation 3.9. Note that this integration is over the whole data distribution $p(X, Y)$ in the com-

plete n-dimensional feature space. In general this joint distribution is unknown. Thus, It is hoped that the training set is a representative sample from this true distribution, but in many situations this might not be the case.

$$\mathbb{E}\{\boldsymbol{Err}\} = \int\int \boldsymbol{Err}(f(X), Y)p(X, Y)dXdY \tag{3.9}$$

Considering the MSE error in equation 3.8, it is possible to decompose the expected Error in three main terms: a bias component; a variance component; and a component associated to the noise in the observations. This is shown in the equation 3.10. A detailed derivation of this can be found in any of the following references (Domingos 2000, Hastie et al. 2003, Bishop 2006).

$$\mathbb{E}\{\boldsymbol{Err}\} = \ (bias)^2 + \ variance + noise \tag{3.10}$$

The first expression is referred to as the square of the *bias*. This gives a measure of the extent to which the average predictions of the learned model differs from the optimal predictions associated to a "real function" underlying the data available. Thus bias measures the systematic loss incurred by a learner. The bias is independent of the training set, and is zero for a classifier that always makes the optimal prediction.

The second term is referred to as the *variance*. This represents the variation of the prediction of learned classifiers, when using different training samples. The variance measures the loss incurred by its fluctuations around the central tendency in response to different training sets. The variance is independent of the true value of the predicted variable, and is zero for a classifier that always makes the same prediction regardless of the training set.

The last term, also called the *irreducible error* (Hastie et al. 2003), is beyond our control and it is independent of the classification model chosen. Note that

this bias-variance decomposition can be made for other types of errors such as the *zero-one loss* (Domingos 2000).

Whilst the complexity of the model selected is increased, the variance tends to increase and the bias tend to decrease, which is also related to the overfitting problem mentioned before. On the contrary, when a more rigid model (not flexible enough to follow all characteristic in the data) is chosen, the bias tend to increase and the variance tend to decrease (Bishop 2006).

This phenomenon can be clearly appreciated in figure 3.3 (which is adapted from (Yoo et al. 2008)), where the error associated to bias and variance are expressed as a function of the model complexity.

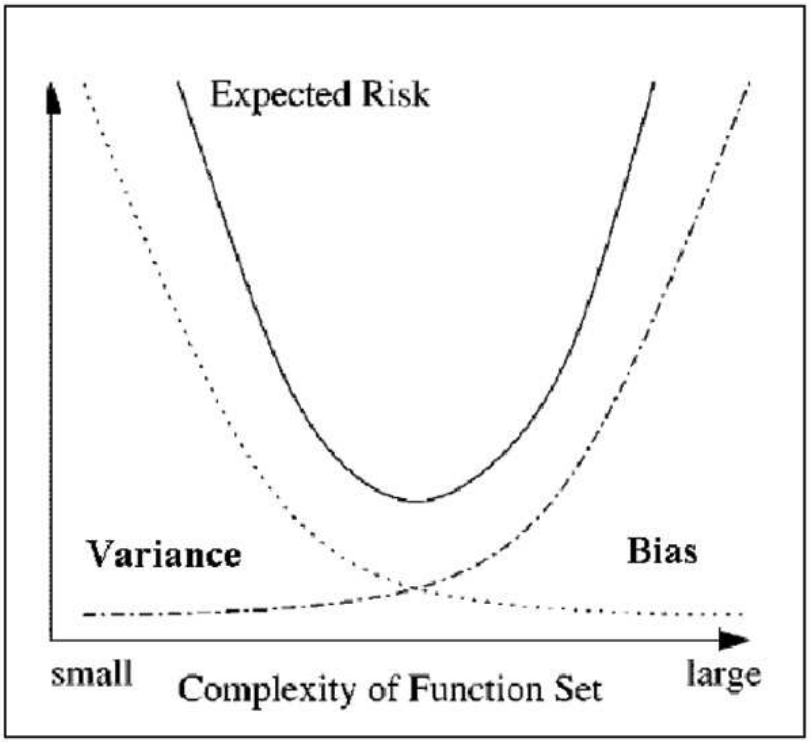


Figure 3.3: Bias-variance trade-off as function of model complexity

The best fitting function $f(X)$ for a given sample is therefore an equilibrium trade-off between the bias and the variance contribution in order to minimize the test error (Hastie et al. 2003). A good fitting function should have both,

a small bias and a small variance. The function should be flexible enough to capture the data, but it should also be simple enough to avoid overfitting.

Summarizing the discussion in this section, for the problem of classification we need to focus our attention in to three main factor (Alpaydin 2004), as follows: the complexity of the hypothesis we fit the data; the amount of training data available; and the generalization error on new examples.

Chapter 4

One-Class Classification for prediction of PPI

4.1 Introduction

In the previous chapter we have discussed potential problems associated with the application of conventional binary classification techniques for the prediction of PPI. These are mainly related to the selection of a trustable set of negative examples (non-interacting protein pairs) and an imbalance class situation. In order to deal with this situation, we have proposed the use of one-class classification (OCC) methods for this task. These methods employ only examples of one class (real interacting protein pairs) to generate a predictive model able to infer new PPI. In addition OCC techniques are able to deal with imbalanced class situations.

In this chapter, we introduce the concept of "One-Class Classification" and how these type of methods can be utilized for the problem of PPI prediction. Several OCC algorithms are described in detail. These OCC methods are then employed for the prediction of PPI based on the integrative learning analysis

of diverse biological data. Furthermore we present the results of a comparative evaluation between OCC and conventional binary classification techniques for this task. We consider different scenarios, for instance varying the number of negative examples used for training purposes. Among different OCC, methods we found that the parzen OCC approach performs competitively with traditional approaches and in many situations outperforms them. Finally, we evaluate the ability of the parzen OCC approach to predict new potential PPI targets, validating these results by searching for biological evidence in the literature.

4.2 One-class Classification

The common issue of OCC problems is that feature information is available for only one of the classes, called the *target class*, and this is employed to generate a classification model. The OCC model is constructed with the aim of characterizing and describing the target examples, and afterwards is used to distinguish target examples from all other examples which have been classified into a single different category called the *outlier class*. The general task in OCC can be regarded as being similar to conventional binary classification methods, in that a decision boundary or separation model is used to separate examples of the two classes (target and outliers). However, OCC methods face a harder task because the decision boundary is mainly supported by examples of the target class and hence less information is employed to build and validate it. Consequently, a sufficiently representative sample of target examples is needed to generate a more accurate descriptive model, in order to improve the OCC performance.

In this research we consider the task of prediction of PPI as an OCC problem, in the sense that only examples of one class (positive interaction examples) are available and/or trustable, becoming the *target class*. The resulting classification model is independent of the kind and quality of the set of negative examples

employed, because the OCC approach is mainly based on the description of examples of the target class; this could potentially solve the problem of trustability associated with the selection of the negative class. In order to develop a comparative performance evaluation between OCC and conventional classification methods, a set of negative examples should be selected as the *outlier class*. This is because it is necessary to use examples of both classes for training and testing purposes when considering conventional binary classifiers. Under these conditions, the performance of OCC methods can be evaluated in a manner similar to the one for conventional binary classification techniques, by estimating the misclassification error, i.e the target class error (or false-negative rate) and, when outlier examples are available, the outlier class error (or false-positive rate).

OCC methods can be classified according to the way in which they analyze, describe and generate a model for the separation of targets and outlier examples (Tax and Duin 2004). Here we consider two types, as follows. *(A) Density estimation* methods based on the estimation of the probability density distribution of the training data using some probabilistic model (e.g. Gaussian distribution). A threshold is selected and then used to compare with the density of new objects in order to classify them. *(B) Boundary* methods based on the generation of a frontier or boundary around the target objects, which is optimized to accept most of the target examples and at the same time reject most of the outliers. Four different OCC learning approaches were evaluated in this research, namely three density estimation methods (single Gaussian estimation, mixture of Gaussian and Parzen density estimation) and a boundary approach (Support vector data description SVDD). The dd_tools Matlab toolbox (`http://www-ict.ewi.tudelft.nl/~davidt/dd_tools.html`) was utilized to develop the experiments associated with the application and evaluation of all OCC methods.

The four OCC methods evaluated in this research (3 density and 1 boundary approaches) are described in the following sections.

4.2.1 Gaussian density estimation

This is the simplest of the OCC density approaches. The examples of the target class used for training are modeled as a Gaussian distribution. In the dd_tools implementation, the complete density estimation is not obtained and only the Mahalanobis distance is employed and calculated for each example X as:

$$f(X) = (X - \mu)^T \Sigma^{-1} (X - \mu) \tag{4.1}$$

where the mean μ and the covariance matrix $\sum$ are estimated from the entire sample of objects used. The $f(X)$ value for new objects is then compared against a threshold θ and classified as a target if $f(X) \leq \theta$ or else as an outlier.

4.2.2 Mixture of Gaussian density estimation

In this case, a linear combination of several (i.e. N) different Gaussian distributions is employed to model the target class examples used for training, obtaining a more flexible model compared with the single Gaussian distribution approach. The training data is divided into N different clusters, each of which is modeled by a single Gaussian distribution. The distance function $f(X)$ changes in this case to the form:

$$f(X) = \sum_{i=1}^{N} \alpha_i \exp(-(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i)) \tag{4.2}$$

where α_i are the mixing coefficients. The parameters of each cluster μ_i, $\sum_i$, and α_i are optimized using the EM algorithm. A threshold θ is fixed again and used to classify new objects as in the previous case. For this approach it is possible

to include outlier objects in the training phase, setting independent mixtures of Gaussian distributions for both target and outlier examples, considering N_{target} and $N_{outlier}$ different clusters. The number of clusters considered for target and outlier data should be fixed and consequently can be varied in order to obtain an optimal performance of the model.

4.2.3 Parzen density estimation

In Parzen density estimation, an independent Gaussian distribution is considered for each one of the T target objects used to train the model. Consequently, in this case the distances to all training objects have to be considered. In the dd_tools implementation of this approach, the function $f(X)$ is as follows:

$$f(X) = \sum_{i=1}^{T} \exp(-(X - X_i)^T h^{-2} (X - X_i)) \tag{4.3}$$

The smoothing parameter h, commonly called the *Parzen width*, is introduced here and is related to the width of a region R (in a Gaussian space) generated around each object in order to separate the target from outlier zones. The rest of the classification process is similar to the previous density approaches. The value of h can be varied in order to optimize the performance of the model.

4.2.4 Support vector data description (SVDD)

This technique is a boundary approach based on the binary Support Vector Machines (SVM) theory. The aim of SVDD is to create a closed hyper-spherically shaped boundary around the target class examples used to train the model. Following the description in (Tax 2001, Tax and Duin 2004) the hyper-sphere is characterized by the centre a and radius R, and is supported for several objects as in the case of SVM. The objective then is to minimize the volume of the sphere,

which is possible by minimizing the value of R^2. This minimization problem is similar to that in the SVM approach and consequently it is possible to generate the same kind of approximation solution. The SVDD method can also employ a more flexible representation of the data using different kernel functions (i.e. linear, polynomial and Gaussian kernels). This approach permits the use of outlier examples in the training stage in order to generate a tighter description of the hyper-spherical boundary. The kernel type and its respective parameters can be varied in this implementation, in order to obtain optimal performance conditions.

4.2.5 Final Remarks

Previously we defined the problem of prediction of PPI and how machine learning techniques are employed for this, considering this task as a classical binary classification problem. The focus of this research is to face several drawbacks related to this binary classification approach. Studying the use of OCC methods as a possible solution to these problems.

The main assumption made in this thesis, is that OCC models will be able to properly characterize the feature space associated to positive PPI examples, and this characterization will be able to correctly discriminate between positive and negative PPI examples. For this, its results crucial to have a sufficient and representative sample of objects of the positive PPI class (Tax and Duin 2004). Additionally it is important to create/generate an adequate set of biological features associated to our problem. These issues will be discussed later in this chapter.

The issues previously discussed in chapter 3 when using conventional or classical binary classification models, such as the definition of the error, atypical training data (not able to characterize the real data), complexity of the solution

(learning model), the curse of dimensionality, the generalization of the method selected, also appear in one-class classification. Some problems become even more prominent.

In conventional classification, data from two classes are available, the decision boundary is supported from both sides by example objects. Most conventional classifiers assume more or less equally balanced data classes and do not work well when one class is severely under sampled or even completely absent. Because in one-class classification only examples of one class of data are available, only one side of the boundary can be determined. It results hard to decide on the basis of just one class how tightly the boundary should fit in each of the directions around the data. it is even harder to decide with features should be used to find the best separation of the target and outlier class. This situation becomes an important disadvantage faced by OCC methods for the specific task of prediction of PPI. However, this also becomes one of the principal objectives of our research. Motivated by the fact that the selection of a negative gold standard set introduce some bias into the learning process (Ben-Hur and Noble 2006).

For the computation of the error function (in equation 3.6), the joint probability $p(X, Y)$ should be known. In the case of one-class classification only the probability of the target class (really interacting proteins pairs) is known. Thus in the definition of the Bayes decision in equation 3.5, we only have information for the conditional probability $p(Y_{k=+1}|X)$. This might introduce extra difficulties when employing OCC methods.

In one-class classification a boundary should be defined in all directions around the data. In particular when the boundary of the data is long and non-convex, the required number of training objects might be very high. So it is to be expected one-class classifiers will require a larger sample size in comparison with conventional ones. Later in this chapter we will evaluate this situation.

The most straightforward method to obtain a one-class classifier is to estimate the probability density of the training data and to set a threshold on this density. Here we will consider density methods such as the Gaussian and the Parzen density. When the training sample data is sufficiently high and a flexible density model is used (for example the Parzen density estimation), this approach works very well. Unfortunately, it requires a large number of training samples to overcome the problem of curse of dimensionality. If the dimensionality of the data and the complexity of the density model is restricted, this can be avoided, but then a large bias might be introduced when the model does not fit the data appropriately.

An important characteristic of one-class classifiers is their robustness against the presence of few outliers in the training data. here we assumed that the training set is a characteristic representation of the target distribution. however it might happen that this training set is already contaminated by some outliers. This situation can also be related to the existence of noise in the training data. Although an OCC method should accept as much as possible objects from the target set, these outliers should still be rejected. Using for instance the threshold for the OCC density methods, some robustness is automatically incorporated (Tax 2001).

The OCC methods described previously have different characteristics concerning the number of parameters W associated that have to be chosen beforehand. When a large number of free parameters is involved, it results difficult to estimate appropriate values for them. These parameters are often called the "magic parameters" because they often have a big influence on the final performance and no clear rules are given how to set them. When these parameters are set correctly, good performance will be achieved, but when they are set incorrectly, the method might completely fail. This is also related to the complexity

of the learning model selected. The number of free parameters should be small to avoid a too flexible model and rapid overfitting to training data. In our case, the Parzen density approach contains the lowest number of free parameters.

In the next sections of this chapter, we will present the results of a comparative evaluation between OCC and conventional binary classification techniques employed for the task of prediction of protein-protein interactions.

4.3 Comparative performance evaluation

4.3.1 Reference data set

In this research we focused on the prediction of co-complexed protein pairs (pairs of proteins which are co-members of the same complex). In order to evaluate different machine learning methods, we need a reference data set (gold standard) containing positive and negative examples. We used the same gold standard sets employed by Lin et al. (Lin et al. 2004) for the study of PPI in yeast. These comprise 2,104 positive examples (true interacting protein pairs) derived from the MIPS complex catalogue (Mewes et al. 2002) and 172,409 negative examples (non-interacting protein pairs) related to protein pairs where the members are localized in different cell compartments and consequently are not likely to interact between them. This reference data set is a subset of the one used by Jansen et al. (Jansen et al. 2003), considering only examples where complete information for each one of the biological features is available.

4.3.2 Biological features

An important motivation for this research is that the integration of diverse kinds of biological data/information could potentially improve our ability to predict protein-protein interactions. Four different types of biological information were

considered following (Jansen et al. 2003) and (Lin et al. 2004):

m-RNA expression, following the assumption that proteins which are members of the same complex are commonly expressed simultaneously. The Pearson correlation was estimated for every protein pair considering two different well known studies: the Rosetta compendium (Hughes et al. 2000) and cell cycle time series analysis (Cho et al. 1998), generating two numeric values between -1 and 1 which are incorporated as features.

functional similarity of protein pairs was estimated from the gene ontology (GO) (Ashburner et al. 2000) and the MIPS (Mewes et al. 2002) functional catalog, according to the procedure previously employed in (Jansen et al. 2003), obtaining two new numeric features. The assumption here is that proteins in the same complex tend to share similar functions or to participate in the same biological processes.

These sources of information can be thought of as a hierarchical tree of functional classes in the case of MIPS functional catalog, or a directed acyclic graph (DAG) in the case of GO catalog. Each protein describes a "subtree" of the overall hierarchical tree of classes or a subgraph of the DAG in the case of GO catalog. Given two proteins, it is possible to calculate the intersection tree of the two subtrees associated with these proteins (set of functional classes two proteins share). This estimation is then made for the complete list of protein pairs ($\sim$18 million in yeast), and thus a distribution of intersection trees is obtained. The "functional similarity" between two proteins is finally defined as the frequency at which the intersection tree of the two proteins occurs in the distribution. The more specific the shared functional annotation is, the smaller is the functional similarity frequency.

Essentiality information (Mewes et al. 2002), assuming that two proteins in the same complex are essential or non essential for cell survival. This feature is then characterized by three possible categories (i.e. both proteins are essential or both are non-essential or only one of them is essential), and is represented by a three dimensional vector taking discrete values of +1 or -1 according to each case.

High-throughput experimental interaction data from Y2H and mass spectrometry based experiments were integrated as features. Four different experimental studies have been considered (Uetz et al. 2000, Ito et al. 2001, Gavin et al. 2002, Ho et al. 2002). In each case, a discrete value of +1 or -1 is assigned to indicate whether the components of a protein pair do interact or do not interact respectively.

Numerical features were normalized to obtain a distribution with a mean of 0 and standard deviation of 1, in order to to put all data in the same range of values and to avoid possible numerical difficulties associated with imbalanced ranges. Every pair of proteins available in the reference data set was represented by a 11-dimensional vector X_i containing the information for the biological features considered here, and a label Y_i which can take two values depending on whether each of the proteins pairs do really interact ($Y_i = 1$) or not ($Y_i = -1$).

4.3.3 Conventional Machine Learning Methods

A representative group of conventional or traditional machine learning techniques, which have been previously used for the task of PPI prediction, was selected in order to undertake a comparative performance evaluation with OCC methods for this specific task. These include: Decision Trees (DT), Naive

Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM). The WEKA machine learning library (Witten and Frank 2005) was used to perform the experiments related to DT, NB and LR, while the evaluation of SVM was carried out using the MATLAB interface to the SVM-light toolbox (`http://svmlight.joachims.org`).

4.3.4 Performance evaluation

OCC and conventional learning approaches were evaluated in different training/testing scenarios varying, for instance, the number of negative examples used to train each of the models. A ten-fold cross validation procedure was carried out for every evaluation, in order to assess the performance variability of the models generated. In each situation, the negative examples which were not utilized in the training step were also included in the testing evaluation. This testing strategy differs from previous approaches used for this task, where only a fraction or sub-sample of the negative gold standard examples was considered to test the models. We think that by including all the available putative negative information each time we test our models, we are carrying out a more relevant and at the same time more challenging evaluation for the prediction of PPI.

Several learning methods evaluated here have parameters to be tuned in order to optimize their performance. Including the Parzen density estimation, SVDD, mixture of Gaussian density estimation and SVM. These parameters have to be tuned entirely from the training set (independent data set different to the one employed for testing classification models). For this, a nested (inner or internal) ten-fold cross validation procedure was developed for each of these classifiers. The nested cross validation procedure is a standard method to deal with this situation, helping to reduce bias of models evaluated (Varma and Simon 2006).

Receiver Operator Characteristic (ROC) curves, illustrating the tradeoff be-

tween the false-positive rates and true-positive rates, were generated for each approach under the different scenarios evaluated. The area under the ROC curve (AUC) was calculated for each case to evaluate the overall performance of different learning algorithms. AUC scores seem to be a better evaluation measure than simple accuracy in imbalanced class problems (Huang and Ling 2005).

We also calculated partial AUC scores, which are related to the normalised area under a fraction of the whole ROC curve which represents a condition of special interest. For example, in the situation of severe class imbalance it seems more relevant to evaluate the performance in the region of low values of false-positive rates (Drummond and Holte 2005), which is the case in the prediction of PPI tasks. In our approach we are interested in evaluating and comparing the performance of the different classifiers under conditions of a low false-positive rate. The aim of this is to maximise the number of real interacting protein pairs predicted while minimizing the number of false-positive predicted ones. This is of special interest for biologists working in the identification and validation of new PPI, because they can focus on the study of only the top ranked predicted PPI targets, instead of evaluating many randomly selected protein pairs. We considered the area under the ROC curve up to the first 50 false-positive examples (AUC-50), which has become a commonly accepted performance measure for this specific task (Ben-Hur and Noble 2005, Qi et al. 2006).

Mean values and standard deviation for AUC and AUC-50 were calculated, based on the ten fold cross-validation individual results, in order to compare the performance of different approaches. When the difference was unclear between the AUC or AUC-50 values for two methods, the Wilcoxon signed rank statistical test (Wilcoxon 1945) for the median of the differences between them was computed considering a 5% significance level, in order to obtain stronger

evidence that one of the methods performed better than the other.

4.3.5 Evaluation of diverse OCC methods

Four different OCC methods were used for the problem of PPI prediction including: Gaussian density estimation, Mixture of Gaussian density estimation, Parzen density estimation and Support Vector Data Description (SVDD). The methods were evaluated on a balanced class set using all the positive examples available and an equal size sample of negative examples randomly selected from the whole negative gold standard set. This was done because some of the OCC methods can take advantage of the use of a sample of negative examples to improve their performance. This procedure was repeated ten times using diverse sub-samples of negative pairs. The results of the estimation of AUC and AUC-50 scores for the OCC performance evaluation are shown in Table 4.1 where the mean and standard deviation are given.

Table 4.1: Comparison of AUC and AUC-50 values for different learning methods evaluated

Method	AUC	AUC-50
OCC methods:		
SVDD	0.9766 ± 0.0032	0.2451 ± 0.0321
Gaussian	0.9377 ± 0.0136	0.1224 ± 0.0136
Mixture of Gaussian	0.9855 ± 0.0094	0.2262 ± 0.0515
Parzen	0.9801 ± 0.0075	0.4010 ± 0.0282
Conventional methods:		
Decision trees (DT)	0.9946 ± 0.0033	0.2129 ± 0.1903
Naive Bayes (NB)	0.9908 ± 0.0017	0.2299 ± 0.0275
Logistic Regression (LR)	0.9928 ± 0.0018	0.0917 ± 0.0307
Support Vector Machines (SVM)	0.9934 ± 0.0018	0.2683 ± 0.0251

The results for the global AUC scores show that there is no significant difference between most of the OCC methods evaluated, with the exception of the

simple Gaussian density estimation method which exhibits the lowest overall performance. On the contrary, the analysis of the results for the AUC-50 scores clearly shows that the Parzen density estimation method (AUC-50 = 0.401) by far outperforms the rest of the OCC methods considered here. The good performance obtained by the Parzen method can be explained because this density estimation method takes into account the information of every target example available. This is different to the rest of the OCC approaches evaluated, where for example only an average probability density estimation from the available data is employed, as in the case of Gaussian and Mixture of Gaussian approaches, or in the case of the SVDD method where just a few examples are utilised to support a boundary between target and outlier examples.

The second best performance for OCC methods considering AUC-50 scores is obtained by the SVDD approach using a Gaussian kernel (AUC-50 = 0.2455). We note that a recent paper by Alashwal et al. (Alashwal et al. 2006) used one-class support vector machines (OCSVM) (Schölkopf et al. 2001), which is an extension of the classical binary SVM technique, to deal with the task of prediction of PPI. In that work, the authors only considered one biological feature based on protein sequence and domain information, reporting that the best results are obtained using a Gaussian kernel. In contrast, in our research we evaluated several different OCC approaches, used diverse biological features and also carried out a comparative performance evaluation with several conventional binary classification methods. Moreover, it has been shown that the SVDD and OCSVM techniques give equivalent solutions (Tax and Duin 2004, Schölkopf et al. 2001) when using a Gaussian kernel.

4.3.6 Comparative evaluation between OCC and conventional classifiers

The Parzen OCC method was selected, due to its good performance, to be compared in a more exhaustive evaluation with several conventional classifiers such as Decision Trees (DT), Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM). Firstly, all the learning approaches were evaluated on the same ten different balanced class sets previously used. Estimates for AUC and AUC-50 scores for these experiments are given in Table 4.1.

Comparative analysis of overall AUC scores shows that conventional classifiers perform only slightly better that the Parzen OCC approach. This was expected because the task associated with OCC only uses examples of one class to generate a classification model. However, in relation to the AUC-50 comparative evaluation, we found that the Parzen OCC approach clearly outperforms all conventional classification techniques (AUC-50 = 0.401). The performance of conventional classifiers in these cases is only comparable with some of the other OCC methods previously evaluated, and is sometimes worse as in the case of the LR approach. SVM showed the best performance for the conventional classifiers (AUC-50 = 0.2687). It is interesting to note that DT exhibits high variability compared with the rest of the methods evaluated. The detailed analysis of AUC-50 results shows that in some of the ten fold cross validation subsets DT performs better than OCC methods, but in others (the majority) it performs very poorly. The Wilcoxon signed rank test (Wilcoxon 1945) was applied in this case demonstrating that the Parzen OCC method effectively outperforms the rest of conventional classifiers.

The difference between the AUC and AUC-50 analysis can be clearly appreciated from the ROC curves of the different learning methods evaluated (see Figure 4.1).

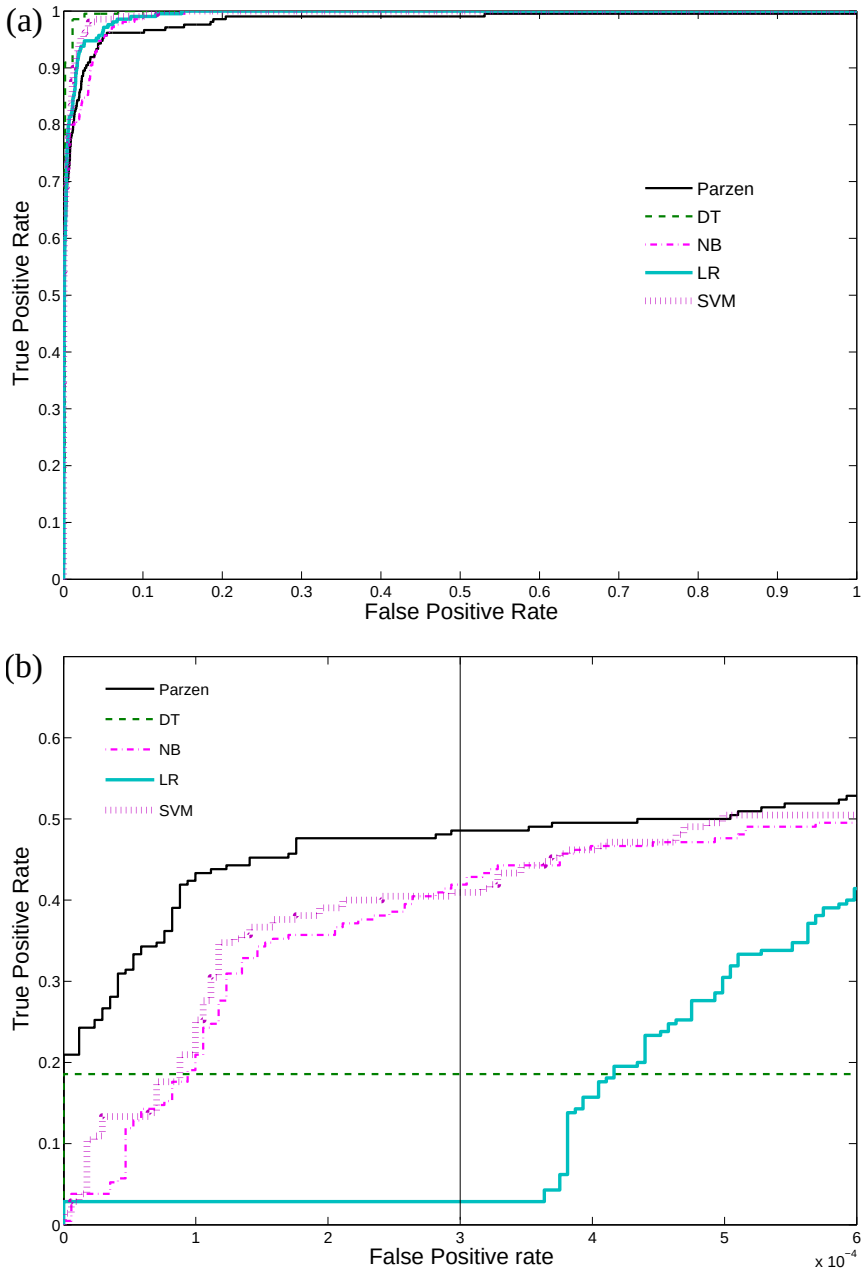


Figure 4.1: Example of ROC curve analysis: (a) Whole ROC curves for the different learning methods evaluated. (b) Partial ROC curves for the different learning methods evaluated. The vertical line indicates the point where approximately the first 50 false-positive examples are reached.

Figure 4.1(a) shows an example of the ROC curves for the different learning techniques used in the evaluation of one cross validation subset. No important differences between these ROC curves is observed and consequently there is no

significant difference in total AUC scores. When we focus on the portion of these curves related to the AUC-50 region, presented in Figure 4.1(b), there are clear differences in the performance of the diverse methods. In this region, the Parzen OCC method outperformed the rest of the conventional learning approaches evaluated. This is still the case if we extend the partial AUC analysis up to the first 100 false-positive examples. This corroborates our assumption that analysis based on partial AUC scores (i.e. AUC-50) is more appropriate than that using overall AUC scores, for predicting PPI.

4.4 Evaluation of different scenarios

4.4.1 Comparative evaluation on different scenarios

We also evaluated and compared the effect of the use of negative examples in the performance of the diverse learning approaches. Different scenarios were generated varying the number of negative examples used for training the respective models, from none to all of the negative examples available. Figure 4.2 shows the performance results, measured as AUC-50 scores, for all the situations considered.

Firstly we analysed the cases where less negative than positive examples were used to train the models, including the balanced class scenario when 2,104 negative examples are employed. The Parzen OCC method clearly outperforms the rest of conventional learning techniques, exhibiting a very stable (almost invariant) performance in the different situations. This can be explained because it only uses positive examples for training purposes. On the contrary, the performance of most of conventional classification methods, with the exception of NB, tends to decrease as less negative information is used. SVM exhibits the best performance for binary classifiers followed by the NB approach. DT and LR

exhibit low performance and high variability compared with the rest of methods evaluated. Note that in the situation where no negative examples are used, only the Parzen OCC method can be employed and consequently no results for conventional classifiers are available.

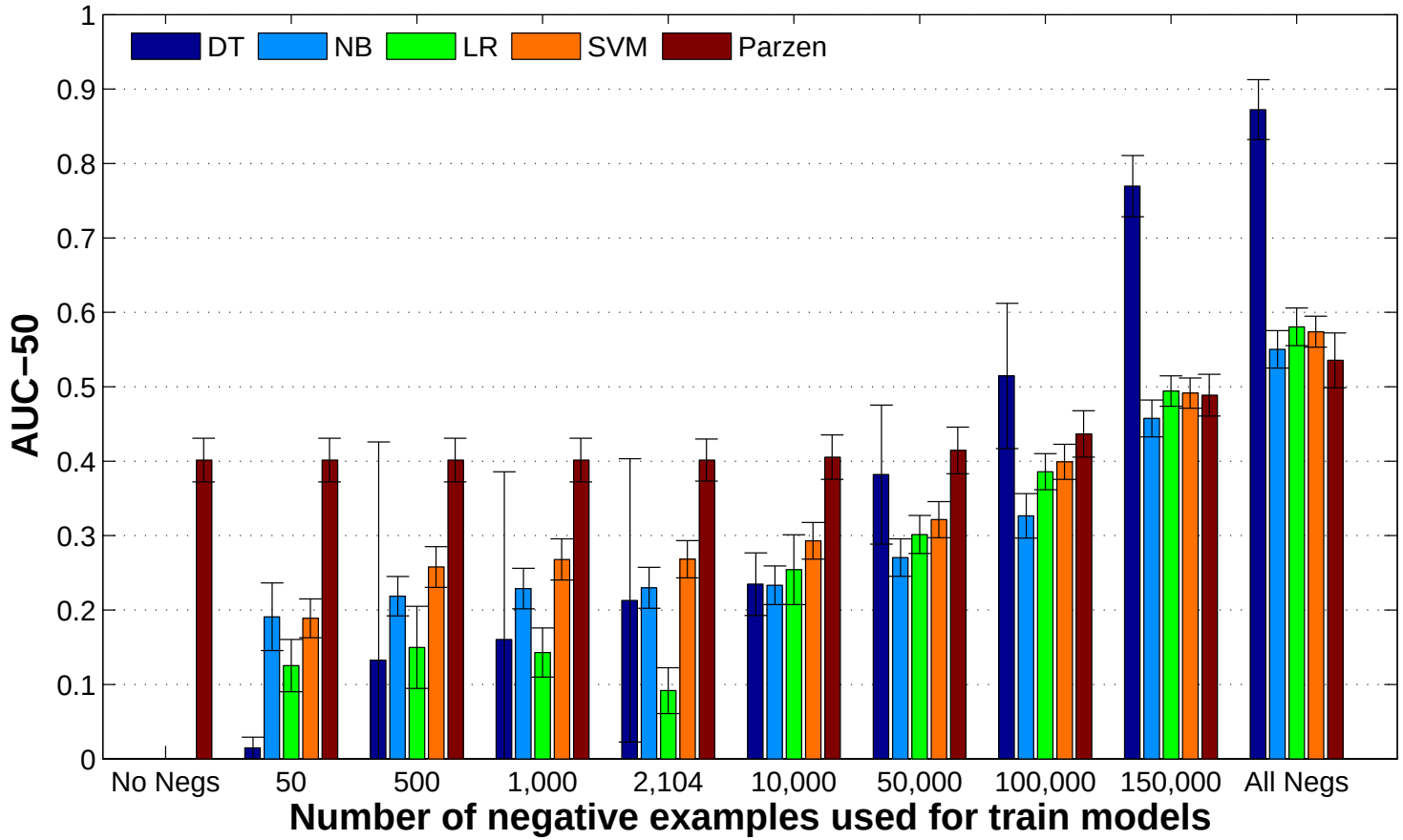


Figure 4.2: AUC-50 comparison for different learning methods evaluated, showing the effect of reducing and incrementing the number of negative examples used to train the models. The balanced class scenario is when 2,104 negative examples are used for training. Note that no corrective action was taken for any of the imbalanced class situations.

The effect in the performance of conventional classifiers suggests that classification models generated by these type of techniques, are not able to correctly discriminate between positive from negative examples. The presence of negative examples seems to be very important for the quality of these classifiers. Parzen OCC models, based only on positive examples information, are more affective to discriminate between both classes under these conditions.

The analysis is quite different for scenarios where more negative than positive examples are employed to train the models. The Parzen density estimation OCC technique tends to maintain its performance stability and a significant increment in the AUC-50 performance only occurs when more than 50,000 negative examples are employed. This can be explained because in these cases the models were tested on a reduced number of negative examples (most of the negative information is used to train the models). The performance of conventional classifiers tends to increase gradually as more negative examples are incorporated for the generation of their respective classification models. This was expected because these techniques can take advantage of the negative object class information.

These results suggest that performance of conventional classification techniques is strongly influenced by the presence of negative examples information. Only under under these conditions, models generated by these kind of techniques seems to be able to correctly discriminate between positive and negative examples. Interestingly we also noted that complexity of conventional classification models tends to increase under these scenarios, (i.e the size of DT models tend to increase as more negative examples are employed in the training phase; more examples are employed as support objects in SVM models). This also support our previous suggestion about the effect of negative examples information on conventional classification models. This results to be the main difference with Parzen OCC models, which does not employ negative information to generate a

predictive model, and consequently maintains its complexity unaffected through the different scenarios evaluated

The Parzen OCC method performs very competitively in most of the scenarios evaluated, and outperforms the other methods up to the case where 50,000 negative examples are used for training. At this point, the DT technique performs as well as the Parzen OCC approach. Thereafter, the DT method outperforms all the rest of the learning approaches, suggesting that DT is the traditional binary learning approach most influenced by the availability of the negative class information. Other conventional classifiers evaluated (NB, LR and SVM) do not exhibit outstanding performance and slightly outperform the Parzen OCC method only when all available negative examples are used.

Finally, we studied the effect of imbalanced classes on the performance of the different classifiers. While OCC methods are intrinsically able to cope with this situation, this is not the case for conventional classifiers. Consequently, some strategy is needed to deal with the imbalanced class problem. Here we used a cost-sensitive analysis, where the misclassification cost for examples of the minority class is bigger than the misclassification cost for the majority class (note that on the different scenarios the minority class is not always the same).

In situations where fewer negative than positive examples were used, we observed an increment in the performance of most of the conventional classifiers, reaching AUC-50 scores similar to those obtained for each approach in the balanced class scenario. The exception is the NB approach, the performance of which was almost invariant in these cases.

When more negative than positive examples were used, the AUC-50 performance for all conventional classifiers tended to decrease in comparison with those obtained without cost-sensitive analysis. This can be explained because in these cases the classification model is generated considering positive and nega-

tive examples information in a balanced way and is not biased towards negative class information. Another accepted strategy to deal with the imbalanced class problem is to under-sample the majority class; we have done this when training on ten different balanced class sets (see previous section).

The analysis of the results presented in this section strongly suggests that the performance of conventional binary classification models is highly affected by the presence or absence of negative examples. This can also explain the high performance (AUC-50) observed for conventional classifiers when all negative examples are employed for training. Another explanation for this observed high performance is the availability of a high-quality negative gold standard set (protein pairs located on different cell localization), which has been previously discussed in (Ben-Hur and Noble 2005) and (Ben-Hur and Noble 2006). However this will not be the case when undertaking the prediction of PPI on other organisms when protein cell localization information is unavailable.

4.4.2 Comparative evaluation when less biological information is available

Most of the previous studies which use a machine learning integrative approach for PPI prediction have been developed considering yeast as a model organism, mainly because more types of biological information are available and consequently it is possible to use these to obtain new evidence and insights about this problem. For other organisms, where less information is accessible, the problem of the inference of PPI is more complex and difficult. In our research we addressed this possible scenario by developing another comparative performance evaluation between OCC and conventional classifiers, considering the case where a reduced number of biological features is available. Two biological features were extracted from the original data set used here, the GO and MIPS

functional annotations, which have been identified to play an important role in the PPI prediction task for yeast in several previous studies (Lin et al. 2004, Lu et al. 2005, Ben-Hur and Noble 2005, Qi et al. 2006). Consequently, by removing these, it is possible to generate a more difficult classification task.

Similarly to the previous analysis developed on the complete data set, a comparative performance evaluation of different OCC methods was performed first for this new data-reduced scenario (using the same ten balanced class sets as before but reducing the number of biological features employed). The AUC-50 scores results for these learning approaches are presented in Figure 4.3, exhibiting again that Parzen density estimation clearly outperforms the rest of the single OCC techniques evaluated. It is important to note that in this reduced-data scenario, the AUC-50 performance score for all the methods evaluated was drastically reduced. For example in the case of the Parzen OCC method, the AUC-50 scores are reduced from approximately 0.4 when the original complete data set was used to arround 0.2 in the reduced data situation, confirming our assumption that this new scenario represents a more difficult prediction task.

The performance evaluation between selected OCC and conventional learning approaches was also carried out on these data-reduced conditions. Figure 4.3 presents the results of the AUC-50 scores for these conditions. Similarly to the results obtained on the complete-data scenario, both OCC methods (Parzen and OCC combination) outperform the rest of the conventional classifiers evaluated here. The SVM and NB approaches show the best performance for the conventional learning approaches obtaining AUC-50 scores slightly over 10%. For these conditions, the performance of the DT method shows an even more high variability compared with the complete-data scenario, confirming the suspicion that DT are highly dependent on the presence of negative examples and even more so in a reduced-information problem.

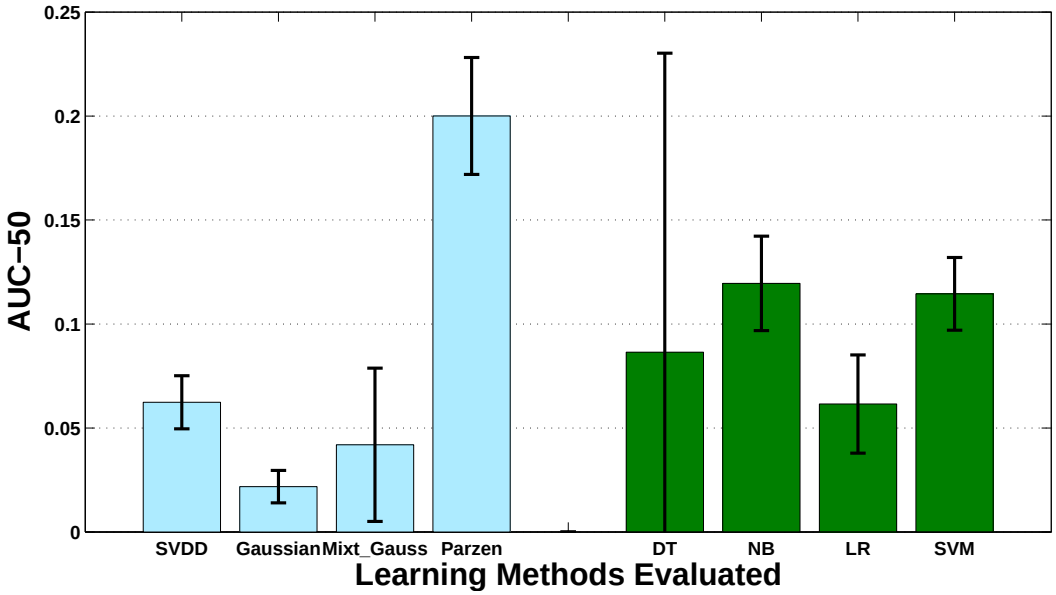


Figure 4.3: AUC-50 comparison for the different learning approaches evaluated in the case where reduced biological information is available. Lights bars present the results for OCC methods: SVDD, Gaussian, mixture of Gaussian and Parzen. Dark bars present the results for conventional classifiers employed: Decision Trees (DT), Naive Bayes (NB), Logistic Regression (LR) and Support Vector Machines (SVM)

4.5 Evaluation of biological feature importance

We then evaluated the individual effect of the different biological features used in this research on the performance of the Parzen OCC approach. For this we removed each of the biological attributes one at time from the data set and tested the effect of this action on the AUC and AUC-50 scores, compared with those obtained when all available biological information is used. Table 4.2 shows the results of this procedure.

The major effect on the Parzen OCC performance occurs when either functional similarity or m-RNA expression data are removed. This is consistent with results previously reported in the literature (Lin et al. 2004, Lu et al. 2005, Qi et al. 2006). It is interesting to observe that the overall AUC performance only increases when high-throughput information is removed, which can be explained

due to the high false-positive and false-negative rates associated with these kinds of features.

Table 4.2: Evaluation of the individual effect of the different biological attributes in the performance of the OCC parzen approach

Feature description	AUC	AUC-50
ALL features	0.9801 ± 0.0075	0.4010 ± 0.0282
GO removed	0.9186 ± 0.0121	0.2094 ± 0.0189
MIPS removed	0.9412 ± 0.0135	0.1983 ± 0.0225
m-RNA expression removed	0.9775 ± 0.0050	0.1883 ± 0.0238
Essentiality removed	0.9800 ± 0.0081	0.3380 ± 0.0273
High-throughput removed	0.9887 ± 0.0037	0.3463 ± 0.0261

4.6 Prediction of new potential PPI targets using Parzen OCC method

Finally, we evaluated the ability of the Parzen OCC approach to predict new potential PPI, which could be used as a targets in future investigations. For this we generated a new set of random protein pairs which were not included in our positive and negative gold standards sets. We were able to collect a set of approximately 518,000 protein pair examples with complete biological information from the data previously used in (Jansen et al. 2003). We classified the examples in the random set using the Parzen OCC model trained on all positive examples available (parameters being optimized on ten fold cross validation procedure), and found that 928 of them were predicted as a new potential PPI.

We focused on the analysis of the top 50 new potential PPI with the highest prediction scores generated by the Parzen OCC model. This score is the probability associated with the positive examples class and consequently can be seen as a confidence value. To validate our predictions we employed the

IntAct database (Kerrien et al. 2007) (`http://www.ebi.ac.uk/intact`), which compiles molecular interactions reported in published literature, containing information for around 50,000 binary protein interactions for yeast (May 2006). We found that of the 50 top ranked examples, 36 were supported by at least one reference in INTACT. These were mostly associated with mass spectrometry experiments which are related to the identification of groups of proteins that interact to form complexes. This is statistically significant considering that if we randomly selected 50 protein pairs not in the positive gold standard, the probability that 36 of them will be annotated in INTACT is very low ($p < 10^{-77}$) using Fisher's exact test (Fisher 1922). The list of the top 50 potential new PPI targets predicted by the Parzen OCC model is given in the Table 4.3.

4.7 Conclusions

The research described in this chapter has focused on the application and evaluation of one-class classification (OCC) methods for the problem of prediction of protein-protein interaction (PPI). We also considered the use of diverse biological data types in order to develop a joint integrative learning analysis.

Among various OCC methods evaluated, the Parzen OCC density estimation approach clearly exhibited the best performance. This can be explained because the Parzen OCC technique utilises all examples in the training set to generate a classification model unlike the other OCC methods investigated here. This approach was then selected to develop a comparative performance evaluation against several well known conventional machine learning methods. Different scenarios were considered varying the number of negative examples used to train the models. We found that the Parzen OCC approach performs very competitively and outperforms the rest of conventional classifiers in most of the situations, up to the case where the ratio of negative to positive examples is

Table 4.3: List of 50 highly ranked new potential PPI targets predicted by the Parzen OCC method

No	ID-1	ID-2	P
1	YDR025W	YLR029C	0.93420
2	YOL039W	YOL139C	0.93085
3	YBR189W	YOR063W	0.92811
4	YKL156W	YPL131W	0.92766
5	YBR118W	YPL131W	0.92748
6	YML063W	YPL131W	0.92665
7	YKL156W	YPL143W	0.92578
8	YGL135W	YNL178W	0.92496
9	YBR048W	YLR029C	0.92441
10	YHR010W	YNL178W	0.92375
11	YBR189W	YPL143W	0.92253
12	YBL092W	YML063W	0.92183
13	YBR189W	YPL237W	0.92140
14	YEL034W	YOL127W	0.91935
15	YBR189W	YMR260C	0.91858
16	YEL034W	YLR340W	0.91823
17	YDL082W	YNL178W	0.91801
18	YDR382W	YNL178W	0.91736
19	YBR118W	YLR029C	0.91652
20	YKL156W	YNL244C	0.91650
21	YML063W	YNL244C	0.91649
22	YDL136W	YNL244C	0.91628
23	YDL082W	YLR249W	0.91615
24	YGL135W	YHL015W	0.91591
25	YEL034W	YOR063W	0.91568
26	YDL191W	YNL244C	0.91460
27	YDR064W	YGL135W	0.91417
28	YEL034W	YKL060C	0.91353
29	YHL015W	YPL220W	0.91250
30	YBR118W	YOL040C	0.91235
31	YHR010W	YNL244C	0.91169
32	YDR025W	YLR249W	0.91152
33	YBR118W	YOL127W	0.91072
34	YML024W	YNL244C	0.91037
35	YNL244C	YPL220W	0.90981
36	YML024W	YPL143W	0.90814
37	YHR010W	YPL237W	0.90769
38	YER131W	YPL143W	0.90761
39	YER131W	YLR075W	0.90642
40	YBL027W	YHL015W	0.90638
41	YKL156W	YOR063W	0.90594
42	YGL135W	YOL139C	0.90561
43	YBR189W	YLR075W	0.90505
44	YDL130W	YDR064W	0.90369
45	YDL136W	YPL237W	0.90351
46	YDL136W	YNL178W	0.90273
47	YAL003W	YDR418W	0.90200
48	YEL034W	YOL040C	0.90180
49	YDL082W	YOL040C	0.90125
50	YDR385W	YOL040C	0.90094

approximately 25 to 1.

We have demonstrated that for this specific task, the performance of conventional binary classification approaches is highly influenced by the quantity of negative examples used to train the respective models. This suggests that classification models generated from these type of methods are more reliant on negative information (in this case an untrustworthy set of negative PPI examples) than on positive information (experimentally corroborated PPI examples).

Our results indicate that the task of the prediction of PPI can indeed be formulated as an OCC problem where the predictive model is based on real (trustworthy) PPI data. In the specific case of prediction of co-complexed proteins, we found that the Parzen OCC method is able to generate models which perform competitively with those generated by conventional classifiers, independently of the quality and quantity of the negative examples available. We have also carried out an initial study on the ability of the Parzen OCC approach to predict new potential PPI targets, showing that many of the highly ranked new predictions can be validated by reference to published results in the literature. Most of the work associated to this chapter has been included in a referred publication in (Reyes and Gilbert 2007).

4.8 Summary

In this chapter we focused on the application and evaluation of OCC methods for the prediction of PPI, considering the use of diverse biological data types in order to develop a joint integrative learning analysis. Among various OCC methods evaluated, the Parzen OCC approach exhibited the best performance. We then developed a comparative performance evaluation between Parzen OCC and several conventional machine learning methods. The Parzen OCC approach performs competitively and outperforms the other conventional classifiers in many

scenarios evaluated. We have demonstrated that the performance of conventional classifiers is highly influenced by the quantity of negative examples used to train the respective models. These models are more reliant on negative information (untrustworthy set of non-interacting protein pairs) than on positive information (experimentally corroborated PPI examples).

In the next chapter we will evaluate another potential drawback related to the task of prediction of PPI, this time associated with the composition of the positive gold standard set (set of real interacting protein pairs).

Chapter 5

Prediction of non-Ribosomal PPI

5.1 Introduction

In this chapter, we discuss and evaluate the effect of a new potential drawback for the problem of prediction of PPI. Positive gold standard sets frequently employed for the task of prediction of co-complexed PPI contain a high proportion of instances related to ribosomal proteins. We demonstrate that this situation biases the classification results and additionally that the prediction of non-ribosomal based PPI is a much more difficult task.

In order to improve the performance of this subtask, we implement two strategies: integration of more biological data into the classification process, including data from mRNA expression experiments and protein secondary structure information, and investigating several strategies for combining diverse one-class classification (OCC) models generated from different subsets of biological data.

We demonstrate that the integration of new biological data has a positive effect on the performance of the Parzen OCC classifier, especially in the case of secondary structure information. In relation to the combination strategy evaluation, the results indicate that the weighted average combination approach

exhibits the best results, significantly improving the performance attained by any single classification model evaluated.

5.2 Analysis of Positive Gold Standard Set composition

In order to predict co-complex protein pairs (pairs of proteins which are co-members of the same protein complex) we need a reference data set or gold standard set containing positive (true interacting protein pairs) and negative examples (non-interacting protein pairs). Although only positive examples are needed in order to train OCC methods, a set of negative ones is still required to obtain a comparable performance evaluation measure. Here we extend the data set we previously employed in (Reyes and Gilbert 2007) to consider a larger number of positive and negative examples. For this we removed the information related to high-throughput experiments for direct detection of PPI. We followed the work in (Jansen et al. 2003) to derive the positive gold standard set from the MIPS complex catalogue (Mewes et al. 2002), and also the negative gold standard set which is related to protein pairs which are present in different cell localizations and consequently are more likely not to interact. A similar reference data set has been employed before in (Jansen et al. 2003, Lin et al. 2004, Lu et al. 2005, Browne et al. 2006). The final data set we employed in this research includes ∼6,700 positive examples and ∼550,000 negative ones, considering only examples where complete information for each one of the biological features were available.

three different types of biological data were employed as features to develop our classification approach, as follows:

- *mRNA expression:* the Pearson correlation is estimated for every protein

pair considering two different studies the Rosetta compendium (Hughes et al. 2000) and cell cycle time series analysis (Cho et al. 1998).

- *Functional similarity* of protein pairs was estimated from the gene ontology (GO) (Ashburner et al. 2000) and the MIPS (Mewes et al. 2002) functional catalog, obtaining two new numeric features. The assumption here is that proteins in the same complex tend to participate in the same biological processes.

- *Essentiality information,* was also used (Mewes et al. 2002), assuming that is more expected that two proteins in the same complex are both essential or non essential but not a mixture of these two attributes.

Analyzing the composition of the positive gold standard set, we found that a high proportion of these examples ($\sim$66%) are related to ribosomal protein pairs. This is because ribosomal protein complexes (cytoplasmic and mitochondrial) are the most numerous among all the different complexes included in the MIPS complex catalogue (Mewes et al. 2002) which contain a large number of proteins. In this research, we argue that this situation could considerably affect the performance of the classifiers, biasing the classifiers to mostly recognize interactions related to ribosomal proteins. In order to assess this situation we proceeded to divide our positive gold standard set in two subsets containing all ribosomal related PPI and all non-ribosomal related PPI respectively, generating at the same time two new classification subtasks related to the prediction of ribosomal and non-ribosomal PPI. The new positive gold standard sets contain $\sim$4,600 and $\sim$2,100 protein pairs respectively. We employed the same negative reference data set in both cases.

The performance of the Parzen OCC approach was evaluated for the three situations considered above. Here we follow the same approach as in the previ-

ous chapter (Chapter 4) to evaluate the performance of different classification methods. Focusing in the performance evaluation of models under conditions of low false-positive rate, aiming to maximize the number of real interacting protein pairs predicted while minimizing the number of false-positive predicted ones. The performance of the Parzen OCC classifier for the different tasks mentioned above is shown in Table 5.1. As in previous chapter, several conventional classifiers were also included in this evaluation (Decision Trees, Support Vector Machines and Naive Bayes). The performance of conventional classifiers is also given in Table 5.1.

Table 5.1: Performance of different classifiers measured as AUC-50 scores. Three cases are evaluated: prediction considering all PPI in the positive gold standard set, prediction of ribosomal PPI and prediction of non-ribosomal PPI. AUC-50 scores given as mean value and standard deviation (in brackets) based on a ten fold cross validation procedure

Classifier	All PPI	ribosomal	non-ribosomal
Parzen OCC	0.5425 (0.0228)	0.7422 (0.0121)	0.1239 (0.0179)
Binary classifiers:			
Decision Trees	0.4916 (0.2902)	0.4808 (0.4486)	0.0439 (0.0280)
Naive Bayes	0.0064 (0.0021)	0.4710 (0.0202)	0.0207 (0.0105)
Support Vector Machines	0.2687 (0.0250)	0.5479 (0.1217)	0.0433 (0.0124)

We observed a clear difference between the performance in the prediction of ribosomal and non-ribosomal PPI. In the case of prediction of ribosomal PPI, the Parzen OCC approach exhibits a high performance of $\sim$0.75 measured as an AUC-50 score. The prediction of non-ribosomal PPI seems to be a more difficult task; here the performance of Parzen OCC approach is significantly reduced to only $\sim$0.12 measured as an AUC-50 score. Interestingly, the performance in the situation when all PPI available in the positive gold standard are employed reach an AUC-50 score of $\sim$0.54 which is in-between the performance of both subtasks. The same behavior was observed when conventional classifiers were evaluated.

The results show that the Parzen OCC approach clearly outperforms all conventional classification techniques for the different tasks evaluated, confirming our previous results reported in (Reyes and Gilbert 2007).

These results suggest that the performance obtained using the whole positive gold standard set is biased towards the prediction of ribosomal related PPI. The high performance exhibited in the prediction of the ribosomal PPI can be explained because they share common patterns in most of the biological features employed in the classification process, specifically those associated with functional similarity and mRNA expression based features. This is not the case when predicting non-ribosomal PPI which appears to be a much more difficult challenge and needs more attention by the scientific community in order to improve its performance. However, a similar positive gold standard set derived from MIPS complex catalogue (Mewes et al. 2002) has been employed in many studies related to the prediction of co-complex PPI (Jansen et al. 2003, Lin et al. 2004, Zhang et al. 2004, Lu et al. 2005, Qi et al. 2006, Browne et al. 2006, Van Berlo et al. 2007). The problem associated with the high proportion of ribosomal related proteins has not been previously reported or addressed according to the best of our knowledge. Furthermore in this chapter we have focused on the task of prediction of non-ribosomal PPI and how to improve the performance of the Parzen OCC method for this task.

5.3 Integration of Biological Information

5.3.1 mRNA Expression Integration

In order to improve the performance of prediction of non-ribosomal PPI, we evaluated the effect of integrating more biological information into the classification process. The first approach developed was related to the integration of informa-

tion associated with mRNA expression experiments. Here we explore the idea that m-RNA expression data obtained under different experimental conditions could give insights about different sets of new potential PPI. This is related to the identification of PPI sub-networks associated with cell adaptation to changing environments proposed and discussed in detail in (Guo et al. 2007). We integrated the data generated in (Gasch et al. 2000) related to yeast stress response. mRNA data previously employed in our study was related to yeast cell-cycle time series analysis (Cho et al. 1998) and the Rosetta compendium (Hughes et al. 2000) which was related to gene mutations and chemical treatments. We evaluated the performance of the Parzen OCC method for this new data set following the same procedure as described in chapter 4. Initially, we considered the case when all the biological features are integrated in a single data set which is then employed to generate and evaluate the performance of the Parzen OCC method, in order to evaluate the individual effect of the different mRNA expression data in the performance of the Parzen classifier. We also considered cases where information related to only one of the mRNA expression experiments is employed. Finally, we considered the situation where no mRNA expression data is employed. The results for all these situations are exhibited in Table 5.2 (middle column).

We observed that when all the data is employed together, the performance of the Parzen OCC classifier is only slightly improved, reaching an AUC-50 score of $\sim$0.14 (compared with an AUC-50 score of 0.1239 in the original situation as shown in Table 5.1). When data from only one mRNA expression experiment is employed we found a significant increment in the performance of the Parzen OCC method for the case of cell-cycle and stress response condition and a slight increment when using Rosetta experiments. The fact that models based on individual mRNA information perform better than the case when all

Table 5.2: Performance for diverse sets of biological data measured as AUC-50 scores. AUC-50 scores are given as mean value and standard deviation (in brackets) based on a ten fold cross validation procedure

Description of data employed	mRNA Integration AUC-50	Plus SS Integration AUC-50
All mRNA expression data	0.1404 (0.0033)	0.2271 (0.0183)
Only Rosetta Compendium	0.1424 (0.0249)	0.2395 (0.0177)
Only Cell-Cycle	0.1859 (0.0208)	0.2344 (0.0146)
Only Stress response	0.2493 (0.0283)	0.2694 (0.0181)
No mRNA expression data	0.1249 (0.0220)	0.2656 (0.0238)

data is integrated together suggests that the integration of features related to diverse mRNA expression conditions does not have a synergistic effect on the performance of the Parzen OCC method. On the contrary, the integration of these features in a single data set seems to induce some kind of misclassification effect and consequently tends to reduce the overall performance. One possible explanation for this situation is that individual mRNA expression data sets (related to different experimental conditions) give different insights to the prediction problem. Moreover, the classifier based on all the features together is not able to correctly discriminate between these situations. Finally the case when no mRNA information is employed exhibits a performance similar to the one obtained in the original situation described in section 5.2. Considering all these results we believe that it might be useful to investigate other ways to combine the information related to individual mRNA predictive models – see section 5.4.

5.3.2 Protein Secondary Structure Integration

Following the idea of integrating more biological information, we investigated the use of protein secondary structure (SS) information. SS information has been employed in recent years for the characterization of protein-protein binding

sites (Neuvirth et al. n.d., Hoskins et al. 2006, Guharoy and Chakrabarti 2007a, Zhou and Qin 2007). However, these approaches consider only a reduced number of PPI which have been crystallized and are available in the Protein Data Bank (PDB) and additionally are focused exclusively on the interaction site region. In our approach we extend this idea to incorporate a larger number of PPI. To the best of our knowledge, this is the first investigation associated with the use of secondary structure information for the prediction of PPI in a broad context.

In order to develop our approach instead of using 3D structure information, we employed the whole linear protein sequence which is available for all yeast proteins. For each protein involved in our study, we predicted the SS and relative solvent accessibility (RSA) for each residue employing the SSPRO program (Cheng et al. 2005). In this case SS, is related to three possible types for each residue: helix (H), strand (E) and the rest(C). RSA is associated with buried (b) or exposed (e) residues. Once SS and RSA sequences have been predicted we faced the problem of how to generate features that could reflect some kind of relationship between SS and RSA for any two proteins. These features were then integrated into our general task of prediction of PPI and so were estimated for each instance included in the positive and negative PPI gold standard sets. A total of 13 features were generated as follows:

- *SS similarity:* Three features were generated based in the similarity of two SS sequences. Local and global alignments scores were estimated using the SSEA software (Fontana et al. 2005). Additionally, we incorporated the common Edit Distance between them.

- *SS and RSA composition:* Four features were generated based on the SS and RSA composition following the work in (Cheng and Baldi 2006). For every protein, a composition vector H,E,C,b,e containing the fraction of each residue type in the whole sequence, was estimated. Then, four sim-

ilarity scores were calculated using dot product, cosine, Gaussian kernel and correlation between any two composition vectors.

- *Ratios:* Six features were generated based on the ratios of the composition of SS and RSA (measured this time as the number of residues of each type) and the total protein sequence length.

Firstly, we evaluated the performance of the Parzen OCC method when only the 13 features based on SS and RSA information were employed. However, the results (AUC-50 scores) in this case were very poor (results not reported here). Further we evaluated the effect of integrating these 13 features with the rest of the biological data previously employed. For this we used the same data set previously evaluated in section 5.3.1, incorporating the SS and RSA information for each of them. The results related to the performance of the Parzen OCC approach when secondary information is integrated are shown in table 5.2 (left column).

We could see that the integration of secondary structure information has the effect of significantly incrementing the performance of the Parzen OCC approach in all situations (different subsets of biological data). This suggests that this type of information can indeed contribute to improving the performance of PPI prediction. Even though each of these features do not perform well when employed alone, it seems that integration with other types of biological data helps in the discrimination between positive and negative examples in the AUC-50 region. Similar to the analysis developed in section 5.3.1, we again observed that models based on individual mRNA expression conditions perform better than when all biological information is employed together. This confirms our initial assumption that no synergistic effect is obtained when different mRNA expression data is utilized together. However, in this case, the effect seems to be less significant, which can be attributed to the presence of SS features.

Interestingly, the strongest increment in performance is shown in the case when no mRNA expression data is employed at all, more than doubling the performance of the original case. This suggests that the Parzen OCC model generated in this last configuration can give different insights to the problem of prediction of non-ribosomal PPI than those models based on individual mRNA expression information. This is also supported by the fact that SS based features contribute to improving the performance of every model based on individual mRNA data.

5.4 Combination of OCC Models

Based on the results obtained in the previous section, we further investigated the possibility of combining the predictions of different Parzen OCCC models in order to improve the performance of the prediction of non-ribosomal PPI. This exploits the idea of combining models that give us different insights to the problem of prediction of non-ribosomal PPI. Four models evaluated in sections 5.3.1 and 5.3.2 were selected which could potentially satisfy this assumption. Three were based on individual mRNA expression experiments (without SS features) and one was based on SS features with no mRNA information.

5.4.1 Diversity of Classification Models

By combining the predictions of different classifiers we aimed to improve the performance of the overall classification task (Dietterich 2000a). This general approach is known under different names in the literature: *classifier ensembles, ensemble learning systems, mixture of experts, etc.* Other works (Kuncheva and Whitaker 2003, Tsymbal et al. 2005, Tang et al. 2006) have shown that a good ensemble is only possible when the base classifiers perform diversely. This means

correctly classifying and/or misclassifying different sets of objects. However, diversity between classifiers can not ensure that there is an improvement in the overall performance. Without diversity there is no point in investigating the combination of diverse classification models.

In order to evaluate the diversity of the four selected classification methods, we considered three general diversity measures commonly employed in the related literature: *Disagreement measure*, related to the degree of disagreement between two classifiers simply calculating the number of cases where one classifier is correct and the other is incorrect (Ho 1998); *Q statistics*, related in this case to the degree of similarity in the performance between two classifiers (Yule 1900); and *Kohavi-Wolpert variance*, which is associated with the variance derived from the decomposition formula of the classification error of a classifier (Kohavi and Wolpert 1996). To calculate these diversity measures for the four models selected in our approach, we followed the general guidelines proposed in (Kuncheva and Whitaker 2003). In our approach, we are interested in the diversity of different classification models specifically in the AUC-50 region (low false-positive rate values). Thus we adapted the diversity measures as follows: we considered exclusively the first"N" instances with the highest prediction confidence for each of the four Parzen OCC classifiers. We then generated a unique list of instances integrating all selected sets. Finally, instead of considering if an object is correctly or incorrectly classified by a classification model, we focused on whether any object belonged or not to the highest confidence list of each model.

Estimates of these diversity measures are shown in table 5.3. The results are given as mean value and standard deviation (in brackets) based on 10 fold cross validation (10FCV) procedure. These results were estimated using N=150; this value was selected arbitrarily considering that on each evaluation related to the 10FCV procedure around 200 positive examples are classified (non-ribosomal

PPI gold standard set contains a total of ∼2,100 instances). Diversity estimates employing N equals 100 and 200 were also calculated (results not included here) exhibiting similar values. In the case of the Disagreement measure and Q statistics, the average over all binary combinations of the four models selected was calculated. The table 5.3 also shows the theoretical minimum and maximum values for each diversity measure considering the case when four models are combined. The Q statistic measure was normalized to have values between 0 and 1 (maximum diversity) following the approach in (Tsymbal et al. 2005).

Table 5.3: Variability of diverse models employed for combination process

Diversity measure	mean value	Min.	Max.
Disagreement	0.4946 (0.005)	0	1
Q statistic	0.5850 (0.022)	0	1
Kohavi-Wolpert variance	0.1855 (0.002)	0	0.25

From the results in Table 5.3, it is possible to see that the four Parzen OCC classification models selected show a high diversity in all cases. This confirms our initial hypothesis that these models which were induced from diverse biological subsets of data give different insights into the problem of prediction of non-ribosomal PPI. We then are interested in to evaluate if by combining their prediction scores, it might be possible to improve the performance of the overall task.

5.4.2 Combination Strategies

In order to combine the predictions of the four Parzen OCC methods selected, we investigated several strategies commonly employed in the literature. Each classifier in our ensemble assigns a predictive (or confidence) value to every object classified. These individual predictions were then combined in several ways in

order to generate a single prediction score, which is employed for the final classification of diverse instances included in the test set. Four fixed combination rules were firstly investigated, which are related to the Mean, Median, Maximum and Product combination of the predictions of different classifiers. These approaches are fixed in the sense that it is not necessary to optimize any extra parameter(s). Additionally, we investigated the weighted average combination approach, where different weights are assigned to each classifier prediction, and the finally prediction score was calculated by a linear combination of them (Kuncheva 2004).

In order to optimize the performance obtained by the weighted average combination approach (AUC-50 score), we developed the following procedure. Firstly, we constrained the sum of all weights to be equal to 1 (no negative weights were considered). Then we evaluated the performance (AUC-50 score) under different situations assigning different sets of weights to each classifier. For this we considered the whole range of possibilities, varying the weights assigned to each classifier between 0 and 1. Finally, we selected the set of weights exhibiting the highest AUC-50 score. The results derived using these combination strategies are shown in Table 5.4.

Table 5.4: Performance for diverse combination strategies measured as AUC-50 scores. AUC-50 scores given as mean value and standard deviation (in brackets) based on a ten-fold cross validation procedure

Model combination strategy	AUC-50
Mean combination	0.2897 (0.0218)
Median combination	0.2679 (0.0213)
Max combination	0.2226 (0.0234)
Product combination	0.3594 (0.0303)
Weighted average combination	0.3809 (0.0314)

We can see that most of the combination strategies produce an increment in the performance of the prediction of non-ribosomal PPI (with the exemption of

the Maximum rule combination strategy), compared to the performances previously given in Table 5.2. The best performance was obtained when employing the weighted average combination approach. In this case, an AUC-50 score of over 0.38 was achieved, representing a significant increment in the performance of this task. The weights assigned to each classifier in the weighted average combination approach can be assigned a certain degree of importance. In the optimum situation achieved here, the Parzen OCC model based on SS data without mRNA expression information was given the highest weight ($\sim$0.5), followed by the models based on mRNA expression associated with Stress response ($\sim$0.3), cell-cycle($\sim$0.15) and Rosetta compendium($\sim$0.05). The second best performance was achieved by the product combination approach with an AUC-50 score of $\sim$0.36; interestingly, this combination technique seems to perform well if the outcomes of individual classifiers are independent (Duin 2002).

5.5 Conclusions

The research described in this section addressed the problem of the prediction of co-complex PPI using the Parzen OCC method and integrating diverse kinds of biological data. The positive gold standard set usually employed in this task contains a high proportion of ribosomal PPI. We have demonstrated that this situation introduces a bias in the classification task. We also showed that the subtask associated with the prediction of non-ribosomal PPI is a more difficult problem. This subtask has not received attention in the past, and according to the best of our knowledge, our work is the first attempt to deal with this situation.

We focused our efforts on improving the prediction of non-ribosomal PPI. We investigated the effect of integrating new biological information into the process, based on data from mRNA expression experiments and protein secondary struc-

ture (SS) information. We have demonstrated that the integration of data from diverse mRNA expression experiments into a single data set has a negative effect on the performance of the Parzen OCC approach. There is no synergy effect in this case, and Parzen OCC models based on individual mRNA expression experiment outperform the one which integrates all the data. On the other hand, the integration of protein secondary structure information results in a positive effect on the increment of performance of this predictive task. The performance of all of the models evaluated is improved when SS-based features are incorporated into the classification process, including the case when no mRNA expression data is used. These results are very promising, and according to the best of our knowledge this is the first attempt to integrate this kind of information for the prediction of PPI.

Finally, we investigated several strategies to combine predictions of different Parzen OCC models induced from diverse subsets of biological data. Four models were selected for this procedure, three based on individual mRNA expression experiments (without SS information) and one based on SS information (without mRNA expression data). These models exhibited a high degree of diversity in their predictions, corroborating our assumption. We have demonstrated that it is possible to significantly improve the performance of the prediction of non-ribosomal PPI by combining the predictions of several Parzen OCC models. The weighted average combination approach exhibited the best performance, and also gave some insights regarding the relative importance of the different classifiers employed. Most of the work associated to this chapter has been included in a referred publication in (Reyes and Gilbert 2008).

5.6 Summary

In this chapter we evaluated the effect of positive gold standard set composition in the performance of Parzen OCC for the prediction of PPI. We demonstrated that this situation introduces a bias in the classification task. We also showed that the subtask associated with the prediction of non-ribosomal PPI is a more difficult problem.

Focusing our efforts on improving the prediction of non-ribosomal PPI. We investigated the effect of integrating new biological information into the process. We have demonstrated that the integration of data from diverse mRNA expression experiments into a single data set has a negative effect on the performance of the Parzen OCC approach. Parzen OCC models based on individual mRNA expression experiment outperform the one which integrates all the data together. The integration of protein secondary structure information results in a positive effect on the increment of performance for this predictive task. The performance of all of the models evaluated is improved when SS-based features are incorporated into the classification process.

Finally, we investigated several strategies to combine predictions of different Parzen OCC models induced from diverse subsets of biological data. We have demonstrated that it is possible to significantly improve the performance of the prediction of non-ribosomal PPI by combining the predictions of several Parzen OCC models. The weighted average combination approach exhibited the best performance, and also gave some insights regarding the relative importance of the different classifiers employed.

Chapter 6

Analysis and Validation of New Predicted PPI

6.1 Introduction

In previous chapters we have demonstrated that parzen OCC approach can deal efficiently with the problem of prediction of PPI. In this chapter we will present a preliminary evaluation analysis of the capability of the parzen OCC approach to predict new potential PPI targets. For this we generate a new data set of PPI consisting of random protein pairs not contained in previous gold standard sets. We then apply the parzen OCC model to this random set to predict new potential PPI among them. These new PPI can be used afterwards as new, and hopefully more trustable, targets for biologist developing small scale experiments. Here we focus our analysis in three main areas:

- Firstly we consider the new predictions as a PPI network. In this case the proteins are nodes and interactions are indirect edges. We then analyze in detail the topological properties of this network such as power law vertex degree distribution and small world effect.

- Secondly we look for highly connected modules of proteins in the predicted network. For this we employed two clustering techniques commonly applied to this task, clique percolation and Molecular Complex detection methods.

- Finally we look for evidence in the related biological literature and data bases to validate these new predictions.

6.2 Identification of New PPI Targets

In order to develop a preliminary validation study of the ability of the parzen OCC approach to predict new potential PPI targets. We have generated a new set of random protein pairs which were not included in the positive or negative gold standards sets. For this, we follow the same approach as described in section 4.6 of this thesis, collecting in this case a set of ∼1,500,000 protein pair examples with complete biological information from the data previously used in (Jansen et al. 2003).

In Chapter 5, we demonstrated that it is possible to significantly improve the performance of the prediction of non-ribosomal PPI by combining the predictions of several Parzen OCC models, derived form diverse sets of biological information. In this case we follow the same strategy in order to classify the examples in the random set previously mentioned. Firstly, we generated four Parzen OCC models utilizing diverse biological data sets. These models were trained on all positive examples available, employing the same parameters optimized before. These models were employed to make predictions associated to each example in the random PPI set previously mentioned. We then proceeded to combine these predictions in the same way as stated previously in section 5.4 utilizing a weighted average combination methodology. The final combination

approach assigned predictive scores to each example contained in the random PPI set.

We then were able to select a group of new potential PPI targets for further validation. To do this we focused again in the region associated to the AUC-50 region of our combined OCC classification model. We chose a cut-off prediction score obtained when training on all positive examples and testing on the examples of the negative gold standard set, but in this case utilizing a "leave one out cross validation" approach. Negative examples were related to protein pairs which are present in different cell localizations and consequently are more likely not to interact. Those protein pairs which exhibited a prediction score over the selected cut-off were classified as new potential targets.

Finally, a set of 818 new PPI targets, involving a total of 306 different proteins, was predicted using this methodology. Figure 6.1 exhibits the PPI network associated to these 818 new targets predicted using our combination approach. For this we utilized the Cytoscape visualization software (Shannon et al. 2003). The complete list of these new PPI targets is given in Appendix A. Further analysis of the topological properties of this network will be discussed in the next section.

Similarly as was made in section 4.6, here we develop an initial validation analysis of the top ranked new potential PPI with the highest prediction scores. For this, we focused in the analysis of the first 100 new PPI targets and searched for biological evidence in the literature and related databases. In this case we employed the IntAct database (Kerrien et al. 2007), which compiles molecular interactions reported in published literature, containing information for over 50,000 binary protein interactions for yeast. We found that approximately half of these new predictions were supported by at least one reference in these databases. These were mostly associated with mass spectrometry based experiments, which

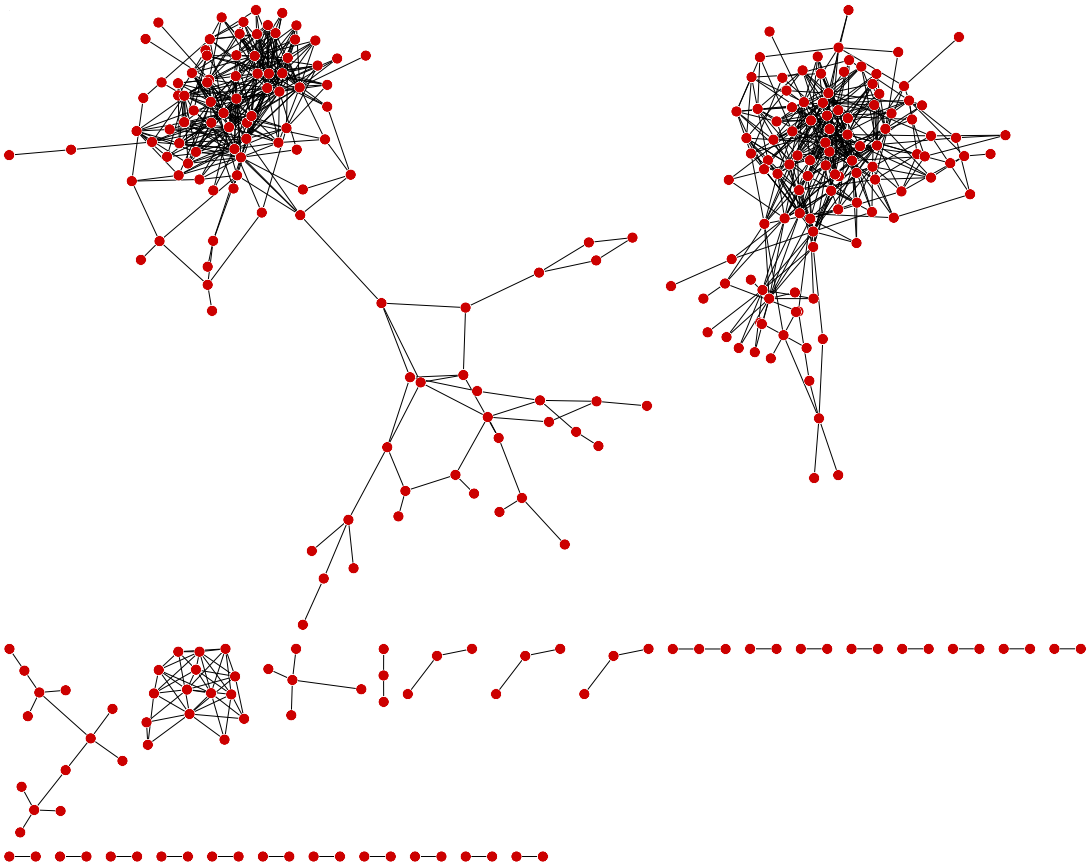


Figure 6.1: Graphical overview of the set of 818 new PPI targets predicted using the combined OCC approach

are related to the identification of groups of proteins that interact to form complexes. This result is statistically significant according to the Fisher's exact test (Fisher 1922). The list of the top 100 potential new PPI targets is exhibited in Table 6.1 for further analysis.

Table 6.1: List of top 100 highly ranked new potential PPI targets predicted by the combination of four Parzen OCC method

No	ID-1	ID-2	No	ID-1	ID-2
1	YDL126C	YDR394W	51	YDL126C	YML092C
2	YDL126C	YDR427W	52	YKL210W	YOR157C
3	YDL126C	YER021W	53	YDL097C	YKL210W
4	YDL126C	YMR314W	54	YDL132W	YKL010C
5	YDL126C	YPR108W	55	YHR200W	YKL210W
6	YDL126C	YOR259C	56	YDR177W	YPR108W
7	YDL126C	YOR157C	57	YER094C	YJR099W
8	YDL126C	YOR261C	58	YJR099W	YOR261C
9	YDL132W	YDL147W	59	YDL132W	YML092C
10	YDR177W	YOR261C	60	YOR124C	YOR249C
11	YER021W	YKL210W	61	YDR394W	YOR249C
12	YER021W	YOR249C	62	YDL132W	YFR050C
13	YDL126C	YOR362C	63	YDL132W	YKL213C
14	YBL041W	YDL126C	64	YGR135W	YOR249C
15	YDL132W	YDR394W	65	YFR004W	YKL210W
16	YDL132W	YER021W	66	YGL048C	YOR124C
17	YDR177W	YER021W	67	YKL022C	YOR117W
18	YDL132W	YPR108W	68	YDL150W	YHR143WA
19	YDR177W	YDR394W	69	YDR092W	YGR253C
20	YDL126C	YGL011C	70	YGR232W	YKL210W
21	YBL041W	YDL132W	71	YDR092W	YJL001W
22	YDL126C	YOR117W	72	YDL147W	YDR177W
23	YDL126C	YFR050C	73	YDR177W	YOR259C
24	YKL210W	YOR117W	74	YKL210W	YKL213C
25	YDL126C	YDL147W	75	YJR099W	YOR117W
26	YDL097C	YDL126C	76	YDL150W	YNR003C
27	YFR050C	YOR249C	77	YJL001W	YKL213C
28	YDL126C	YKL210W	78	YKL145W	YOR124C
29	YKL210W	YPR108W	79	YDL165W	YGR005C
30	YDL126C	YFR004W	80	YKL210W	YOR362C
31	YDR394W	YKL210W	81	YDL147W	YDR092W
32	YDL007W	YDL126C	82	YBL041W	YKL213C
33	YOR157C	YOR249C	83	YHR143WA	YOR174W
34	YOR249C	YOR261C	84	YDL007W	YDR092W
35	YDL132W	YOR261C	85	YDL150W	YOR224C
36	YDR394W	YLL039C	86	YML111W	YPR108W
37	YDL132W	YDR427W	87	YCR093W	YHR143WA
38	YDL132W	YGL011C	88	YDL150W	YKL144C
39	YDL007W	YOR124C	89	YJL197W	YPR108W
40	YDR427W	YOR249C	90	YKL058W	YOR174W
41	YDR394W	YOR124C	91	YDR177W	YOR117W
42	YFR004W	YJR099W	92	YIL075C	YKL010C
43	YBR082C	YDL007W	93	YDR429C	YER025W
44	YGL011C	YOR249C	94	YIL156W	YJL001W
45	YDL126C	YOL038W	95	YDL150W	YOR210W
46	YDL132W	YOR259C	96	YDR092W	YER012W
47	YDR092W	YOR259C	97	YDL126C	YKL010C
48	YDR054C	YOR117W	98	YDL097C	YJL197W
49	YBR082C	YER094C	99	YKR094C	YPR108W
50	YBR082C	YOR261C	100	YBR058C	YDL126C

6.3 Network Topology Analysis of Predicted PPI Network

In previous section we were able to predict a set of 818 new PPI targets. The network of interactions between proteins is usually represented as an interaction graph, "Protein Interaction Network", where proteins are nodes and pairwise interactions are undirect edges. The predicted PPI network previously generated is presented in Figure 6.1.

Graph theory approaches have been applied to describe the topological properties of this kind of networks. The topology of a network refers to the relative connectivity of its nodes, affecting the specific network properties. It has been realized that the architectural features of molecular interaction networks within a cell exhibit similar features to other complex systems, such as the "World Wide Web" or even "social networks". This unexpected similarity indicates that similar laws may govern most complex networks in nature. This enabled the expertise previously acquired in the analysis of large and well-mapped non-biological systems to be employed in the characterization of complicated inter-relationships that govern cellular functions (Barabasi and Oltvai 2004). The relative positions of proteins within the interaction networks might indicate their functional importance. For instance a positive correlation between biological essentiality and graphical connectivity has been demonstrated in (Han et al. 2005).

Considering this, it is important to understand and model the topological and dynamic properties of various biological networks. There are various types of interaction networks in the cell, including protein-protein interaction, metabolic, signalling and transcription-regulatory networks. These biological networks do not work independently, together they form a "network of networks" which is responsible for the behavior of the cell (Barabasi and Oltvai 2004).

There are a number of previous investigations related with the topological analysis of real biological networks, including protein-protein interaction networks (Jeong et al. 2001, Ravasz et al. 2002, Goldberg and Roth 2003, Ravasz and Barabasi 2003, Barabasi and Oltvai 2004, Han et al. 2005, Yook et al. 2004, Li et al. 2006). These analyzes have led to the observation of some apparently recurrent properties of biological networks. The main ones are "power-law distribution", and "small world effect" (Chakrabarti 2005), they are described below.

6.3.1 Power-law distribution

The most elementary characteristic of a graph node is its degree or connectivity. Degree "k" measures how many links a node has to other nodes. In the case of undirected networks as in the PPI graph, "k" is related to the number of edges a node is related to (Barabasi and Oltvai 2004). The degree distribution of this kind of networks is a plot of the count "C_k" of nodes with degree "k", versus the degree "k", typically in a log-log scale (Chakrabarti 2005). Investigations focusing on large real networks have demonstrated that many of them have a scale-free topology, in which the number of nodes follows a power-law distribution (Yook et al. 2004). This means that the number of nodes "C_k" with degree "k" is related to "k" by equation 6.1. Where "c" and "r" are positive constants. The constant "r" is often called the "power-law" exponent. On the contrary, random networks are characterized because most nodes have roughly the same number of links (Yook et al. 2004).

$$C_k = c * k^r \tag{6.1}$$

Recent studies have shown the relevance of this type of connectivity for biological networks. In particular protein-protein interaction networks have been associated to a scale-free topology (Jeong et al. 2001, Giot et al. 2003, Li

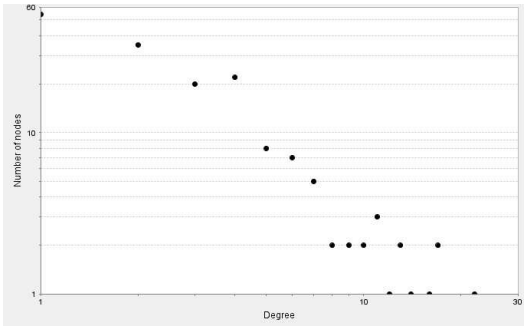
et al. 2004, Li et al. 2006). Scale-free networks are dominated by a few highly connected nodes ("hubs"). These type of networks are resistant to random failure but are vulnerable to targeted attacks, specifically against hubs (Han et al. 2005). This property is related to the robustness of biological networks to perturbations like mutations and environmental stress.

In order to validate the PPI predictions we made using a combination of Parzen OCC models (818 new PPI targets shown in figure 6.1), we estimated the degree distribution of the network associated to these interactions, which is exhibited in figure 6.2. In addition we also estimated the degree distribution for other related PPI networks. Considering the first 300, 500 and 1,500 PPI with the highest prediction scores. This was done to compare the stability of the network topological properties associated to our PPI predictions. Plots for these networks are also exhibited in figure 6.2. These plots were made using the NetworkAnalyzer (Assenov et al. 2008) plugin for the the Cystoscape software (Shannon et al. 2003).

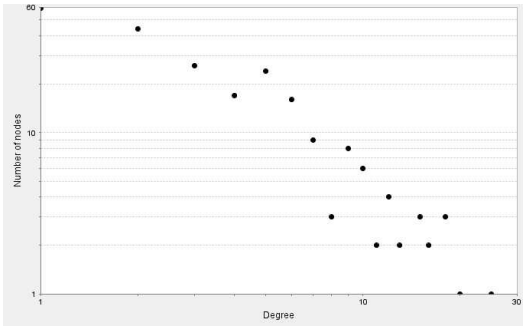
The results of node degree distribution of these PPI networks exhibits that all of them show evidence of a scale-free topology. Each of them follow the power-law degree distribution, indicating that they are all described by scale-free networks. It is important to notice that all networks analysed exhibit a similar degree exponent ($r \sim 1.4$ - 1.5).

6.3.2 Small world effect

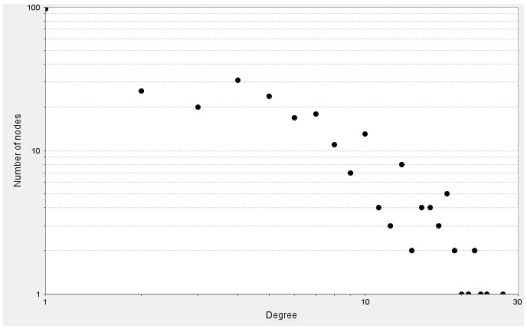
A common feature of many real networks is that any two nodes can be connected through a path of a few links only (Barabasi and Oltvai 2004). This so-called "small-world effect". was originally observed in the research of social networks and is often characterized as the famous "six degrees of separation" (Chakrabarti 2005). Scale-free networks are generally small, their path length is much shorter



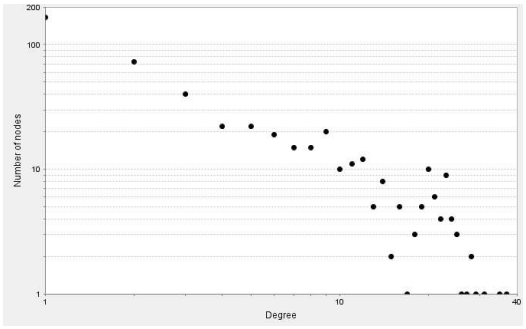
(a) 300 PPI predictions



(b) 500 PPI predictions



(c) 818 PPI predictions,



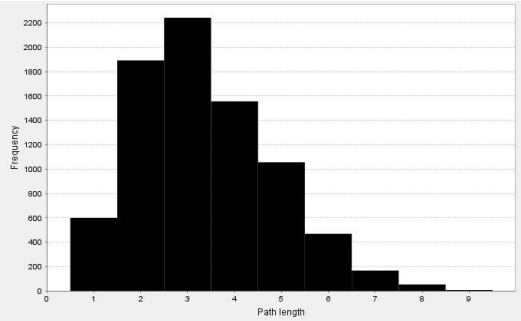
(d) 1,500 PPI predictions

Figure 6.2: Node degree distribution of predicted PPI networks. The plot in (c) exhibits the node degree distribution of the PPI network associated to the 818 new predictions using the classification model based on the combination of various Parzen OCC models (AUC-50 based cut-off). (a), (b) and (d) show the node degree distribution for PPI networks when 300, 500 and 1,500 interactions are included respectively

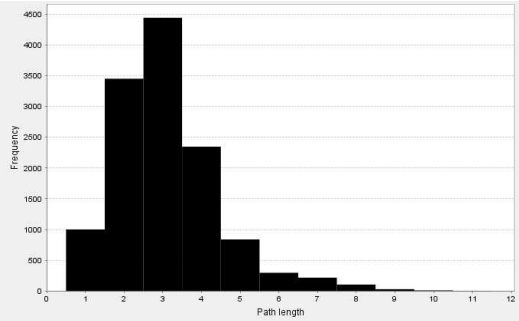
than predicted by the small-world effect for random networks. Within the cell, the "small-world effect" was first observed for metabolism, where paths of only three to four reactions can link most pairs of metabolites (Barabasi and Oltvai 2004). The shortest path length indicates that local perturbations in metabolite concentrations could reach the whole network very quickly. In protein-protein interaction networks, highly connected nodes ("hubs") are not connected to each other and instead connect to proteins with only a few interactions (Barabasi and Oltvai 2004).

Here we analyzed the small-world properties of the predicted PPI network. The distance between any two nodes is defined as the number of edges along the

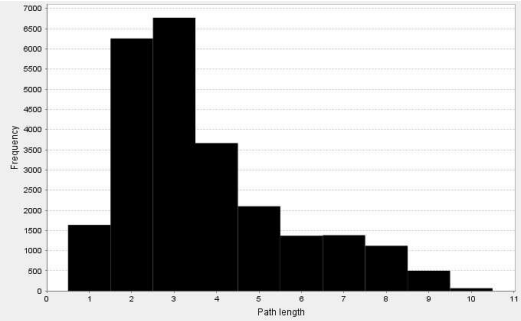
shortest path connecting them. We estimated the shortest path length distribution for the predicted PPI networks previously considered in section 6.1. These results are exhibited in figure 6.3.



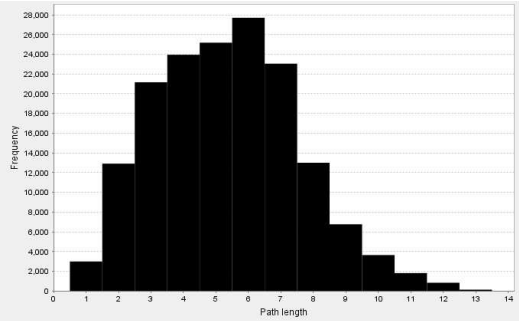
(a) 300 PPI predictions



(b) 500 PPI predictions



(c) 818 PPI predictions,



(d) 1,500 PPI predictions

Figure 6.3: Shortest path length distribution of predicted PPI networks. The plot in (c) exhibits the path length distribution of the PPI network associated to the 818 new predictions using the classification model based on the combination of various Parzen OCC models (AUC-50 based cut-off). (a), (b) and (d) show the the path length distribution for PPI networks when 300, 500 and 1,500 interactions are included respectively

These results indicate that the predicted PPI networks exhibit relatively short paths between their nodes, which can be associated to small-world properties of them. This especially evident in the network associated to the AUC-50 cut-off in figure 6.3.c (818 predited PPI), where most nodes are connected by paths of length between 2-4. The same behavior is observed when only 300 and 500 predicted PPI are included in the analysis (figure 6.3.a and 6.3.b respectively). The path length distribution when 1,500 predicted PPI are utilized

shows a different pattern. In this situation the path lengths are clearly bigger, which can be explained because in this case we are including for the analysis more false-positive predictions (over AUC-50 cut-off region).

6.4 Identification of new PPI Complexes

The network of interactions between proteins is generally represented as an interaction graph, where nodes represent proteins and edges represent pairwise interactions. An important property associated to currently available protein-protein interaction networks is that they tend to be fragmented into many distinct clusters (Yook et al. 2004). These clusters have been usually related to functional modules so-called "protein complexes". The identification of these functional modules from global interaction networks has become one important challenge in systems biology, aiming to understand the relationship between the organization of a network and its function (Bader and Hogue 2003).

To achieve this goal, several clustering methods have been applied to protein interaction networks in order to identify highly connected subgraphs (King et al. 2004, Dunn et al. 2005, Pereira-Leal et al. 2004, Bader and Hogue 2003, Adamcsek et al. 2006, Spirin and Mirny 2003, Rives and Galitski 2003, Arnau et al. 2005, Sharan et al. 2005, Scholtens et al. 2005, Chu et al. 2006). Among these methods, in our research we employed the "Molecular Complex Detection" (MCODE) algorithm (Bader and Hogue 2003). MCODE algorithm utilizes connectivity values in PPI networks to identify complexes, MCODE is focused in the detection of densely connected regions, which is one of the goals associated to our validation analysis. The MCODE algorithm has been successfully employed for this task in several recent investigations (Brohee and van Helden 2006, Zhang et al. 2006).

In order to identify new potential protein complexes, we applied the MCODE

algorithm to our predicted PPI network containing 306 nodes (proteins) and 818 edges (interactions). As a result we were able to identify three independent clusters, which involve a total of 8, 14 and 7 proteins respectively. These correspond to highly connected protein modules, which can potentially be associated to new protein complexes.

Figure 6.4 shows a diagram of these clusters. Proteins involved in each cluster are also listed in tables 6.2, 6.3, 6.4 and 6.5, including details about GO annotations (molecular function and biological process) extracted from the Saccharomyces Genome Database (SGD) (Hong et al. 2008). Following we present a brief description and analysis of each cluster based this information:

Cluster (a):

This cluster groups a set of 8 proteins involved in different RNA translation processes according to the information retrieved from SGD (Hong et al. 2008). Although these proteins have similar cellular function, they are members of different protein complexes. This suggest that our predictive method is able to infer functional relationships between different groups of proteins.

The information related to these new PPI predictions, can be utilized to infer new functional properties of some proteins, improving their functional annotations.

Cluster (b):

This cluster groups a set of 14 proteins. Seven of these proteins are members of known protein complexes related in the RNA transcriptional process at different levels (activation, control and regulation). This again suggests that our predictive model is able to infer functional relationships between different proteins modules. Interestingly, the other half of proteins (7 of 14) are not linked

to any of the protein complexes utilized in the positive gold standard set. These proteins are involved on diverse cell processes such as, metabolic regulation, gene expression regulation and cellular biosynthetic processes.

In this case our predictive approach seems to be able to infer novel relationships between different groups of proteins, which can be investigated and validated in the future by biologists. These results can be used as new evidence for the identification of new members of known protein complexes. Finally the new PPI interactions discovered can also be used to improve functional annotations of several of these proteins.

Cluster (c):

This cluster groups a set of 7 proteins. Although four of these proteins are members of the same protein complex, here we also find two proteins not linked to any of the protein complexes utilized in the positive gold standard set. This corroborates the assumption about the capability of our predictive model to infer novel interactions between different functional groups of proteins. However these new predictions requires further validation. Finally, one protein is involved in a different cellular process, suggesting that our model can potential help in the identification of undiscovered links between different protein complexes.

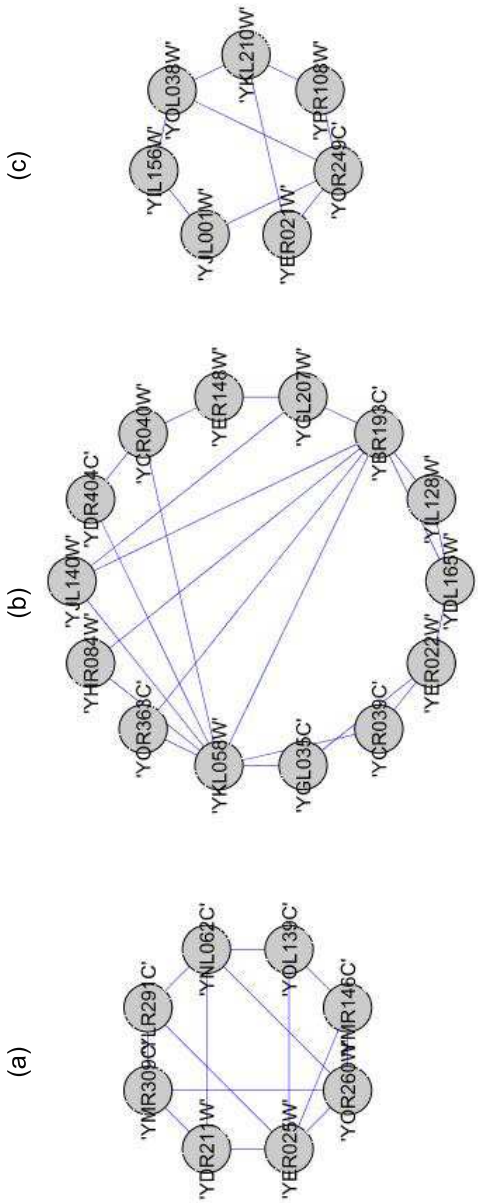


Figure 6.4: Diagram of three clusters discovered employing the MCODE algorithm

Table 6.2: Description of cluster (a) discovered employing the MCODE algorithm.

Protein ID	GO Molecular Function	GO Biological Process
YOR260W	Contributes to guanyl-nucleotide exchange factor activity Translation initiation factor activity	Regulation of translational initiation
YMR309C	Translation initiation factor activity	Translational initiation
YNL062C	Contributes to tRNA activity tRNA binding	translational initiation tRNA methylation
YOL139C	Phosphatidylinositol 3-phosphate binding Translation initiation factor activity	Nuclear-transcribed mRNA catabolic process Regulation of cell cycle Translational initiation
YMR146C	Translation initiation factor activity	Translational initiation
YLR291C	Enzyme regulator activity Contributes to guanyl-nucleotide exchange factor activity Translation initiation factor activity	Regulation of translational initiation
YER025W	Translation initiation factor activity	Translational initiation
YDR211W	guanyl-nucleotide exchange factor activity	Regulation of translational initiation Translation initiation factor activity

Table 6.3: Description of cluster (b) discovered employing the MCODE algorithm.

Protein ID	GO Molecular Function	GO Biological Process
YCR039C	Contributes to DNA bending activity Sequence-specific DNA binding Transcription corepressor activity Sequence-specific DNA binding	Donor selection Regulation of transcription - RNA polymerase II promoter Regulation of transcription
YGL207W	RNA polymerase II transcription elongation factor activity Transcription elongation regulator activity	Nucleosome assembly RNA elongation from RNA polymerase II promoter Transcription initiation from RNA polymerase II promoter
YJL140W	DNA-directed RNA polymerase activity	mRNA export from nucleus Nuclear-transcribed mRNA catabolic process Transcription from RNA polymerase II promoter
YCR040W	DNA bending activity Transcription coactivator activity	Regulation of transcription from RNA polymerase II promoter Regulation of transcription
YIL128W	RNA polymerase II transcription factor activity	methionine metabolic process nucleotide-excision repair Transcription from RNA polymerase II promoter
YKL058W	General RNA polymerase II transcription factor activity	Transcription initiation from RNA polymerase II promoter
YBR193C	RNA polymerase II transcription mediator activity	Transcription from RNA polymerase II promoter
YGL035C	Sequence-specific DNA binding Specific transcriptional repressor activity Sequence-specific DNA binding	Negative regulation of transcription from RNA polymerase II promoter
YDR404C	DNA-directed RNA polymerase activity	Nuclear-transcribed mRNA catabolic process Positive regulation of nuclear-transcribed mRNA Transcription from RNA polymerase II promoter

Table 6.4: Continuation of Table 1.4 for Description of cluster (b) discovered employing the MCODE algorithm.

Protein ID	GO Molecular Function	GO Biological Process
YOR363C	Specific RNA polymerase II transcription factor activity Transcription activator activity	Fatty acid metabolic process Peroxisome organization Positive regulation of transcription Response to oleic acid
YHR084W	Transcription factor activity Sequence-specific DNA binding	Conjugation with cellular fusion Invasive growth in response to glucose limitation Positive regulation of transcription from RNA polymerase II Pseudohyphal growth
YER148W	Chromatin binding DNA bending activity DNA binding General RNA polymerase II transcription factor activity Protein homodimerization activity RNA polymerase I transcription factor activity RNA polymerase III transcription factor activity Sequence-specific DNA binding	General transcription from RNA polymerase II promoter Transcription from RNA polymerase I promoter Transcription from RNA polymerase III promoter transcriptional preinitiation complex assembly
YER022W	NA polymerase II transcription mediator activity	Transcription from RNA polymerase II promoter
YDL165W	Contributes to ubiquitin-protein ligase activity	Negative regulation of transcription from RNA polymerase II promoter Nuclear-transcribed mRNA catabolic process Nuclear-transcribed mRNA poly(A) tail shortening Protein ubiquitination Regulation of cell cycle Regulation of transcription from RNA polymerase II promoter Response to pheromone during conjugation with cellular fusion RNA elongation from RNA polymerase II promoter

Table 6.5: Description of cluster (c) discovered employing the MCODE algorithm.

Protein ID	GO Molecular Function	GO Biological Process
YKL210W	ubiquitin activating enzyme activity	Protein ubiquitination
YER021W	Molecular function unknown	ubiquitin-dependent protein catabolic process
YPR108W	Structural molecule activity	ubiquitin-dependent protein catabolic process
YOL038W	endopeptidase activity	ubiquitin-dependent protein catabolic process
YIL156W	ubiquitin-specific protease activity	Protein deubiquitination
YOR249C	Protein binding ubiquitin-protein ligase activity	Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process Chromatin assembly or disassembly Cyclin catabolic process Mitotic metaphase/anaphase transition Mitotic sister chromatid segregation Mitotic spindle elongation Protein ubiquitination
YJL001W	endopeptidase activity	Ascospore formation Response to stress ubiquitin-dependent protein catabolic process

6.5 Summary

In this chapter, we developed a preliminary study to validate the new potential PPI predicted by using the parzen OCC approach. We focused on the PPI network generated with these predictions, and analyzed different properties related to it. Firstly we focus our analysis in the top 100 ranked PPI predicted, searching in the literature and biological databases for evidence that support them. We then demonstrated that the new predicted PPI interaction network has similar topological properties to those generally observed in most molecular interaction networks. Finally, we discovered three clusters or highly interconnected groups of proteins into the predicted PPI network, and briefly analyzed the biological importance of these novel inferences.

Chapter 7

Prediction of PPI Types

7.1 Introduction

So far we have focused our research in the prediction of PPI, employing several commonly used types of biological information in a proteome-wide scale. Using this data we were able to generate an OCC model to predict new PPI.

Another important source of data is related to structural information. Protein structures are obtained at present by experimental techniques, such as XRay crystallography previously described in Chapter 2. Among the available structures, individual proteins are more frequent than structures of protein complexes. Protein complex structures can be classified according to their life time and binding affinity into four main classes, as obligate permanent interactions involving homo or hetero obligomers and non-obligate transient interactions involving enzyme-inhibitor or non enzyme-inhibitor. In this chapter, we describe a computational approach for the prediction of PPI types employing association rule based classification (ARBC). This includes association rule generation and posterior classification based on the discovered rules. We investigate diverse properties associated with the interface of protein complexes. Aiming to discover

patterns, in the form of association rules, that characterize interaction sites in different PPI types based on these properties.

7.2 Motivation

Protein-Protein Interactions (PPIs) play a key role in many essential biological processes in cells, including signal transduction, transport, cellular motion and gene regulation. The comprehensive analysis of these biological interactions has been regarded as very significant for the understanding of underlying mechanisms involved in cellular processes.

Computational approaches for the prediction of PPI based on atomic level interactions can accurately determine the binding affinity and the specificity of binding partners. Thus, structure based prediction methods including modeling of PPI by homology modeling, threading-based methods and protein-protein docking are more accurate than methods that do not employ structure data. A major drawback of these structure-based methods is the relatively low coverage of available crystallized protein complexes in the Protein Data Bank (PDB) (Berman et al. 2000). This is especially the case for those proteins associated with transient interactions, which is the majority of functional PPIs, and these do not form complexes stable enough for x-ray crystallography (Vakser 2004). Due to these restrictions the detailed analysis of the structure of protein complexes, specifically the area related to the interaction site between proteins, can reveal important clues for the understanding of protein functions and also characterize the specificity of these interaction regions. The prediction of protein interaction sites has gained much attention in recent years with over 20 different methods proposed (Zhou and Qin 2007). Interaction regions can be characterized by a diverse set of physico-chemical properties (Jones and Thornton 1997, William S. J. Valdar 2001, Neuvirth et al. 2004), topo-

logical properties (Davis and Sali n.d.) and conserved residues (Livingstone and Barton 1993). A variety of studies have employed different classification approaches including Support Vector Machines (Bock and Gough 2001, Koike and Takagi 2004, Bradford and Westhead 2004, Zhu et al. 2006) and Random Forests (Chen and Liu 2005). These studies have shown that the interfaces of interaction sites share common properties that distinguish them from the rest of the protein (Chothia and Janin 1975, Jones and Thornton 1996, Jones and Thornton 1997). Despite their good performance in the prediction of protein interaction sites, these machine learning approaches generate final prediction models which do not provide users with explicit rules and thus result in low interpretability of the results and poor knowledge extraction capability.

The identification, analysis and characterization of different PPI types can be classified according to their life time and binding affinity into four main classes (Jones and Thornton 1996, Nooren and Thornton 2003, Bradford and Westhead 2004), as obligate permanent interactions involving homo or hetero obligomers and non-obligate transient interactions involving enzyme-inhibitor or non enzyme-inhibitor. In obligate protein interactions, protomers which are not individually structurally stable *in vivo*, form permanent functional complexes that are stable and exist in their complexed form. Protomers of non-obligate interactions are independently stable and can form transient or permanent complexes. Non enzyme-inhibitors are participants in transient interactions not involving enzymes and their protein inhibitors.

The characterization of PPI types can help for instance in the functional annotation of newly crystallized protein complexes as suggested in (Nooren and Thornton 2003). Several studies have been developed in this direction, focused on the discrimination of different PPI types with the aim of characterizing transient and obligate protein complexes (Nooren and Thornton 2003, Gunasekaran

et al. 2004). These include the statistical analysis of the interface properties (De et al. 2005), and the analysis from an evolutionary view of issues related to these interactions (Mintseris and Weng 2005). A recent computational approach (Zhu et al. 2006) classified binary protein complexes into three categories (obligate interactions, non-obligate interactions and crystal packing) using six interface properties and employing Support Vector Machines (SVM).

In our work we describe a computational approach for the prediction of PPI types employing association rule based classification (ARBC) (Liu et al. 1998, Li et al. 2001), which includes association rule generation and posterior classification based on the discovered rules. In a similar manner to previous approaches we investigate diverse properties associated with the interface of protein complexes. But instead of considering the entire interface area between two proteins we only consider the region associated with domain information by using the SCOP classification (Andreeva et al. 2004). The use of domain profile pairs can provide better prediction of protein interactions than the use of full-length protein sequences as reported in (Wojcik and Schachter 2001). In addition we also incorporate secondary structure information related to these domain-binding sites into our predictive approach. These features appear to be useful for the characterization and classification of binding interfaces as reported recently in (Guharoy and Chakrabarti 2007b). The aim is to discover patterns, in the form of association rules, that characterize interaction sites in different PPI types. An important advantage of using such a classification approach is the interpretability of the final predictive model based on the analysis of the discovered set of rules. Here we focus on the prediction of four different PPI types (i.e. transient enzyme inhibitor/non enzyme inhibitor and permanent homo/hetero obligomers), trying to gain more specific insights into the characterization of diverse kinds of interactions.

7.3 Methods

7.3.1 Interaction Data

We employed the same data set of non-redundant interacting protein complexes reported by (Bradford and Westhead 2004). The set of 147 complexes was selected from a comprehensive set of 180 proteins taken from the PDB. 25 of these 147 complexes are involved in enzyme-inhibitor (ENZ) interactions, 21 in non-enzyme -inhibitor (nonENZ) interactions, 14 in hetero-obligate (HET) interaction, and 87 in homo-obligate (HOM) interactions as shown in Table 7.1. Proteins sharing $> 20\%$ sequence identity with a higher resolution structure of the same complex type were removed. Crystal packing structures were also eliminated by investigating evidence in the literature that the complex occurs naturally and is stable as a dimer. Permanent complexes are more easily available from stable complexes by x-ray crystallography. Transient PPIs often neither form stable complexes nor give good NMR structures. This is reflected in the small number of validated transient complexes available in the PDB.

Table 7.1: Data set of protein complexes

Type Name	Type of Interaction	♯. of Complexes	♯. of Domains
ENZ[a]	enzyme-inhibitors	25	49
nonEnz[b]	non enzyme-inhibitors	21	47
HET[c]	hetero-obligomers	14	33
HOM[d]	homo-obligomers	87	225
Total		**147**	**354**

[a] ENZ: enzyme-inhibitor interactions;
[b] $nonENZ$: non-enzyme-inhibitor interactions;
[c] HET: hetero-obligate interactions;
[d] HOM: homo-obligate interactions.

7.3.2 Definition of *interface* and *dom-face*

An *interface* is a set of interacting atoms whose Solvent Accessible Surface Area (SASA) is decreased by > 1Å upon the formation of a complex (Jones and Thornton 1997). The SASA for each atom was calculated using MSMS (Sanner et al. 1996) with a probe sphere of radius 1.5Å. Given a pair of interacting proteins, we define a set of interacting atoms for a single protomer as a *face*. An *interface* comprises a pair of interacting *faces*. We define the set of atoms comprising the *face* of a single domain as a *dom-face*. In order to calculate *dom-faces*, the interfaces extracted from complexes are mapped onto ranges of SCOP 1.65 domain definitions (Andreeva et al. 2004). A total of 354 SCOP domains were extracted related to form the 147 protein complexes considered in our study of the different PPI types, see Table 7.1.

7.3.3 Description of *dom-face*

We generated 14 different physico-chemical properties and structural features to characterize each of the *dom-faces* considered in our study including: *dom-face* area (df-ASA), hydrophobicity (HH), residue propensity (inPro), number of amino acids (nAA), number of atoms (nAtom), number of Secondary Structure Elements (nSSE), length of consecutive residues (LCS), number of fragments (nFrag), Size ratio of *dom-face* area to domain area (sRatio), Secondary Structure Elements (SSEs) content (Helix, Strand, Non-Regular) and SCOP class number (SCOPClass). Hydrophobicity and residue propensity were analyzed in the same way as Jones and Thornton (Jones and Thornton 1997).

The solvent accessible surface area (SASA) of a *dom-face* is calculated as the sum of the total decreased SASA for the interface atoms in a domain, see Equation 7.1. If A and B are two protomers in the complex AB, $SASA_A$, $SASA_B$ and $SASA_{AB}$ are SASA values for A, B, and AB respectively, and n is the total

number of interface atoms in a domain presented in protomers A and B, then:

$$dom\text{-}face\ Area = (\sum_{i=1}^{n}(SASA_A(i), SASA_B(i)), -SASA_{AB}(i)) \tag{7.1}$$

We employed the hydrophobicity scale of Fauchere and Pliska (Fauchere and Pliska 1983) to estimate the average hydrophobicity value for each *dom-face*. The average hydrophobicity (HH) is calculated using Equation 7.2, where HI_{AA} is the hydrophobicity value for each amino acid residue and N_{AA} is the number of residues in a *dom-face*.

$$HH = \frac{\sum_{i=1}^{l} HI_{AA}}{N_{AA}} \tag{7.2}$$

Residue propensity (inPro) indicates the relative frequency of different amino acid (AA) residues in *dom-faces* of complexes. We estimated residue propensities for all *dom-faces* using Equation 7.3(Jones and Thornton 1996), where AAP_i is the natural logarithm of each AA propensity and N_R is the total number of residues in a *dom-face*.

$$inPro = \frac{\sum_{i=1}^{n} AAP_i}{N_R} \tag{7.3}$$

In order to analyze the size of interaction sites, we computed the ratio between *dom-face* and the whole domain area (SR) employing Equation 7.4.

$$SR = \frac{ASA_{dom\text{-}face}}{ASA_{domain}} \tag{7.4}$$

The sequence continuity in the interaction sites is described by calculating average length (number) of consecutive residues (LCS) and counting the number of consecutive residues (nFrag) in *dom-faces*. The SSE content is calculated by

the percentages of interaction atoms located in Secondary Structure Elements (SSEs), classified using the types defined in DSSP (Kabsch and Sander 1983): helix, strand and non-regular regions (turn, bend and loop). PPI types become the heads of association rules in ARM and the target classes in our classification. We used four different types of PPI, namely enzyme inhibitor/non enzyme inhibitor as transient interaction types and homo/hetero obligomers as permanent interaction types. Other properties estimated for the diverse *dom-face*s analyzed were the SCOP class number at the first level of the SCOP hierarchy, the number of AA, the number of atoms and the number of SSEs present in the different interaction interfaces.

7.3.4 Association Rule Based Classification

The problem of predicting PPI types for a given complex of binary proteins is transformed into the task of assigning a pre-determined target class (i.e., homo/hetero obligate and non-obligate) using properties of interaction sites. We applied an efficient association rules based classification method (ARBC) to perform classification based on rules generated by Association Rule Mining (ARM). Previous studies (Liu et al. 1998, Li et al. 2001) have proposed that ARBC consistently outperforms other rule-based classifiers such as decision trees. ARBC comprises three main steps: association rule generation, pruning association rules and classification based on association rules.

Association rule generation

In our approach we employed Association Rule Mining to discover a set of frequent patterns expressed as association rules describing the relationship between properties of PPI interaction sites and PPI types. Association rules have the form $R : X \rightarrow Y\ [c, s]$, where X and Y are the body and the head of the rule

respectively. X and Y are disjoint predicates ($X \cap Y = \phi$). Each X and Y consists of a conjunction of distinct predicates which describe properties related to interaction sites. Note that we can consider a conjunction as a set for our purposes. In our approach, the heads of all rules Y are restricted to be one of the PPI types considered, which are the target classes defined in this task. The strength of the association rules can be measured in terms of their *support* (s) and *confidence* (c). The support of a rule ($X \rightarrow Y$) is the probability that the cases in a database contain both X and Y. The confidence of the rule is the probability that a case contains Y given that it contains X.

The generation of association rules was carried out employing the Apriori algorithm(Agrawal and Srikant 1994). We used the 10g Oracle Data Miner (ODM) software which implements the Apriori algorithm to compute the type of association rules required for our ARBC approach. We set a minimum support and confidence of 3% and 25% respectively to reduce the number of association rules generated. Association mining is not directly applicable to real valued continuous data such as some of the *dom-face* properties we generated; hence we used discretisation to manipulate continuous attributes before the ARM process was executed. In this process, adjacent values of continuous data were binned into a finite number of intervals.

Pruning association rules

The number of rules generated by ARM can be very large. It is necessary to prune the set of association rules by removing redundant information in order to make the classification more efficient.

Given two rules $R_1 : X_1 \rightarrow Y_1$ and $R_2 : X_2 \rightarrow Y_2$, we define:

Definition 1. **The significance of a rule**: R_1 is more significant than R_2 if and only if either (1) $conf(R_1) > conf(R_2)$ or (2) $conf(R_1) = conf(R_2)$ but

$sup(R_1) > sup(R_2)$ or (3) R_1 has fewer attributes on its left hand side than $R_2\diamond$

Definition 2. **General rule**: Given two rules $R_1 : X_1 \rightarrow Y_1$ and $R_2 : X_2 \rightarrow Y_2$, R_1 is a general rule if and only if $X_1 \subseteq X_2\diamond$

Definition 3. **Overlapping rule**: Given two rules $R_1 : X_1 \rightarrow Y_1$ and $R_2 : X_2 \rightarrow Y_2$, then $R_3 : X_1 \vee X_2 \rightarrow Y_1(conf(R_1), sup(R_1)) \vee Y_2(conf(R_2), sup(R_2))$ is an overlapping rule if and only if $X_1 = X_2$ and $Y_1 \neq Y_2\diamond$

If the body of a rule R_1 is identical to the body of a rule R_2 and the head of rule R_1 is inconsistent with that of rule R_2, then an overlapping rule R_3 between two different PPI types can be identified.

Overlapping rules can be considered as common rules between two or more PPI types. On the other hand unique rules are distinctive patterns which can be used to classify interaction sites into different PPI types. We then evaluated the following condition in order to prune the set of association rules previously generated. Given two rules R_1 and R_2, where R_1 is a general rule R_2, ARBC eliminate R_2 if R_1 has more significance than R_2. Sets of unique and overlapping rules were generated with the pruning procedure used in the classification.

Classification:

In the classification step, we employed the pruned set of unique and overlapping rules to generate a *rule profile* consisting of an $m \times n$ matrix, where m is the number of examples (i.e. *dom-faces*) and n is the number of different association rules obtained after the pruning step. Each row of this matrix represents one of the *dom-faces* considered in our research and is associated with one of the PPI types we wish to classify. The *rule profile* matrix takes values of 1 or 0 depending whether the different rules are contingent or not on the respective *dom-face* example. A similar approach was previously employed in (Viksna et al. 2003) for protein structure comparison. The *rule profile* matrix was generated following

Algorithm 1 and then used as input to the ARBC process.

Algorithm 1 Generation of a rule profile

Input: A set of rules $(R_1, \cdots, R_n)$ and
A set of training data comprising
m objects (O_1, $\cdots$,O_m)

Output: An $m \times n$ matrix, *RProfile*(*i*,*j*)($1 \leq i \leq m$ and $1 \leq j \leq n$)

Method:

1. Sort rules in the descending order of confidence and support

2. for each rule R_j in the descending order of the rules
    for each data object O_i in the training data
        find match between O_i and rule R_j
            if $match(O_i, R_j)$
                set RProfile(i, j) = 1
            else
                set RProfile(i, j) = 0
    end-for
 end-for

We evaluated several classification techniques for this task including Decision Trees (DT), Random Forest (RF), K Nearest Neighbor (KNN), Support

Vector Machines (SVM), and Naive Bayes (NB). The WEKA machine learning library (Witten and Frank 2005) was used to perform these experiments. We also performed conventional classification based only on the physicochemical properties of the different *dom-faces* examples, without generating a set of association rules (CWAR). This was done in order to evaluate if the employment of the ARBC approach could be associated with a loss of information of some interacting complexes due, for example, to the pruning step or the discretisation of continuous value feature information. In all cases, a 10 fold-cross validation procedure was performed. As the task of classification of different PPI types involves imbalanced classes (see table 7.1) we utilized a cost-sensitive strategy, where the misclassification cost for examples of the minority class (PPI types with few examples) is bigger than the misclassification cost for the majority class.

7.4 Results and Discussion

7.4.1 Analysis of *dom-face* Properties

We found that 98.8% of the interaction sites studied are contained within corresponding ranges of SCOP domains. This suggests that the analysis of interaction sites based on structural domains (i.e. *dom-face*) does not lose interaction information.

Average values of diverse *dom-face* properties for different PPI types are shown in Table 7.2. It is possible to observe a distinct difference in the distribution of non-obligate (i.e., ENZ and nonENZ) and obligate (i.e., HET and HOM) complexes. The distribution patterns of *dom-face* area for ENZ are similar to those of nonENZ and the same trend occurs between HET and HOM. In the distribution of the area of interaction sites, obligate PPI types exhibit a greater

variance and in general tend to have larger interaction sites than non-obligate complexes.

The average hydrophobicity (HH) values for ENZ, nonENZ, HET and HOM are respectively 0.40, 0.37, 0.41, and 0.42. Even though average HH values are similar for different PPI types, the distributions of hydrophobicity exhibit distinctive separation patterns between non-obligate and obligate interactions (results not shown here). The distribution of HH for ENZ is similar to nonENZ and that of HET is similar to HOM.

We note that Arg, His, Tyr, Gln and Trp exhibit higher propensities than other amino acids, while Gly has a low propensity in our analysis. Average residue propensities are shown in Table 7.2. HET has the highest residue propensity and HOM the lowest. We also analyzed the top four frequent residues for each interaction type calculating the sum of ASA for each amino acid. Hydrophobic residues including Leu, Ala, and Val frequently occur in types HET and HOM. The charged residue Glu also appears frequently in HET. In nonENZ, charged residues including Asp, Glu, Lys, and Arg are present in the top four frequent residues.

Table 7.2: Average values of the properties

Type	ASA(Å^2)	HH	inPro	nAtom	nAA	nSSE	LCS	nFrag
ENZ	860.42	0.40	0.596	121.73	33.71	11.22	3.3	12.32
nonENZ	823.06	0.37	0.530	106.89	29.59	12.91	2.5	12.91
HET	2237.92	0.41	0.982	344.26	82.56	21.35	3.5	21.35
HOM	1306.37	0.42	0.262	184.55	48.14	13.00	2.9	16.78

ENZ includes not only some polar residues Ser and Tyr but also the charged residue Glu. We observed that the charged residues occur very frequently in all interaction types and appear dominantly in HET. Trp, Cys, and Met rarely occurred in interface area through all types.

We observed that 92% of *dom-faces* are smaller than a half of their domain sizes based on the calculation of ASA values. The average length of consecutive residues (LCS) are 3.3, 2.5, 3.5 and 2.9 for ENZ, nonENZ, HET, and HOM respectively as shown in Table 7.2.

The average distribution of SSE elements (helix, strand and non-regular regions) for different PPI types is shown in Figure 7.1. We have seen that interaction sites are mostly composed of non-regular regions followed by helix and strand regions. ENZ contains 64.15% of non-regular regions, which is the highest percentage. Helix content are greater than 36% in types nonENZ, HET and HOM but are less than 17% in ENZ. Strand content for all types are less than 20% and HET exhibits the lowest value (13.72%).

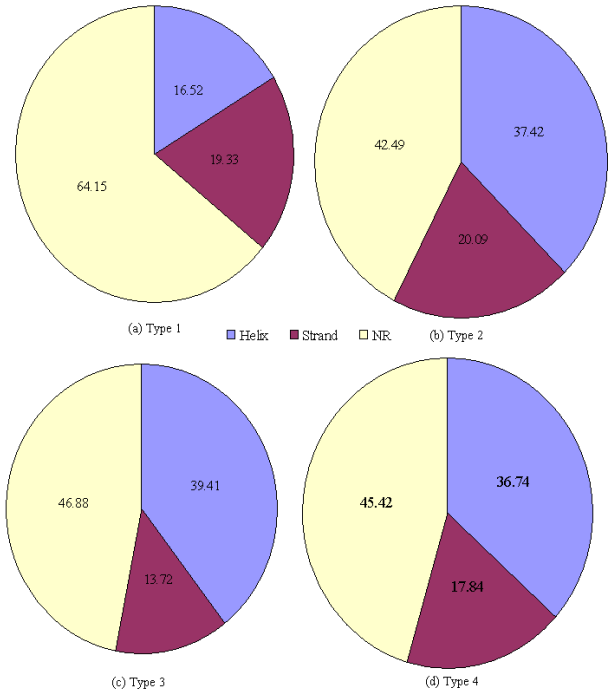


Figure 7.1: average distribution of SSE elements (helix, strand and non-regular regions) for different PPI types

The variation in the number of amino acids (nAA) is similar to that for the number of atoms (nAtom). Average values for nAtom, nAA, nSSE and nFrag are

shown in Table 7.2. We found that values for Types HET and HOM are higher than for Types ENZ and nonENZ in all these categories. The distribution of these properties for ENZ is similar to nonENZ.

7.4.2 Classification of PPI types

We were able to discover a total of 1,168 rules describing associations by employing ARM. After the pruning stage a total of 157 association rules were selected for the classification process. The number of rules associated with types ENZ, nonENZ, HET and HOM are 65, 49, 19, and 24 respectively (Table 7.3). A total of 58 of these are unique, i.e. rules exclusively associated with just one PPI type. The remaining 99 rules are overlapping (non-unique) rules related to two or more PPI types. We are interested in this distinction because unique rules appear to be related to specific characteristics of PPI types, whilst overlapping rules can be related to common attributes of different interaction types or, for instance, to distinctive properties between obligate and non-obligate interactions.

Table 7.3: Number of association rules discovered for each PPI type

Type	♯. of Domains[a]	♯. of Rules[b]	Unique Rules[c]	Overlapping Rules[d]
ENZ	49	65	34 (52.31%)	31 (47.69%)
nonENZ	47	49	16 (32.65%)	33 (67.35%)
HET	33	19	7 (36.84%)	12 (63.16%)
HOM	225	24	1 (4.17%)	23 (95.83%)
Total	354	157	58 (36.94%)	99 (63.06%)

[a]♯ . of Domains: A number of domains in ecah PPI type;
[b]♯ . of Rules: A number of association rules discovered for each PPI type;
[c]Unique Rules: A number of association rules associated with just one PPI type;
[d]Overlapping Rules: A number of rules of which bodies are identical to those of rules in other types.

The performance for different classification methods measured as total accu-

racy over 10 fold cross-validation for ARBC is shown in Table 7.4. Additionally, we performed classification based on the physicochemical properties of the different *dom-faces* (CWAR), and also ARBC classification based on a rule profile generated using only the set of 58 unique rules discovered (UR). Performance results for these approaches are also given in Table 7.4. We have seen that in all these cases SVM exhibited the best performance among diverse classifiers studied, reaching over 99% accuracy in some cases. However this high accuracy suggests that overfitting problems can be associated with the use of SVM. The other classification approaches evaluated still exhibit a high accuracy with the exception of NB. The performance reached by them is comparable to that previously reported in (Zhu et al. 2006) although not exactly the same instances and features were employed. Additionally, we observed that there was no significant appreciable difference between the performance of ARBC and CWAR in most of the situations, although it seems that CWAR performed slightly better than ARBC.

These results strongly suggest that ARBC performs competitively with conventional classification approaches for this task, and consequently the use of ARBC does not involve an important loss of information derived from ARM. The performance of ARBC using only unique rules clearly decreased for all classification methods evaluated, although maintaining an acceptable accuracy of near or over 90% in most of the cases. This suggests that unique rules can be influential in classifying most of the PPI types considered in our study and that overlapping rules are important to improve the accuracy of the classification task. It is important to emphasize that the aim of our research is focused on the advantage of interpretability of the discovered rules, rather than the optimization of the classification task.

We further investigated the influence of SSE information on the classification

Table 7.4: Accuracy for different classification methods

$Method^a$	DT	RF	KNN	SVM	NB
All data[1]*:*					
$ARBC^b$	0.919	0.961	0.941	0.995	0.517
$CWAR^c$	0.922	0.966	0.971	0.999	0.525
UR^d	0.865	0.926	0.888	0.965	0.512
No SSE data[2]*:*					
$ARBC\_WO\_SSE^e$	0.91	0.950	0.931	0.993	0.501
$CWAR\_WO\_SSE^f$	0.917	0.968	0.972	0.986	0.510
$UR\_WO\_SSE^g$	0.800	0.843	0.805	0.889	0.507

[a] $Method$ represents different classification methods such as Decision Tree (DT), Random Forest (RF), K Nearest Neighbor(KNN), Support Vector Machine (SVM) and Naive Bayes (NB);
[b] $ARBC$: Association rule based classification;
[c] $CWAR$: Classification based on physicochemical properties;
[d] UR: ARBC classification using 58 unique association rules;
[e,f,g]: Data sets with exclusion of SSE content from All $data^1$;
[1] All data: Data sets including SSE content;
[2] No SSE data: Data sets without inclusion of SSE content.

of PPI types. We evaluated three different data sets without using the secondary structure elements of proteins, including ARBC_WO_SSE, CWAR_WO_SSE and UR_WO_SSE. Each of the two rule profiles in this case contains a total of only 135 association rules and 43 unique rules. Results for these evaluations are also highlighted in Table 7.4. It was found that in all cases, the performance of diverse classifiers tended to decrease when SSE data was omitted, although only a slightly reduction is observed in most of the classifiers evaluated. Interestingly the major decrement in performance was observed when employing UR_WO_SSE, with accuracies of less than 90% for all classifiers including SVM. These results strongly suggest that SSE content in interaction sites could have an important role in the discrimination of different PPI types for both approaches including ARBC and CWAR.

This implies that the average confidences of the rule sets that include this

SSE content information may be higher than those without it. There were 14.01% (22 out of 157) such rules that included SSE content information and their average confidence was 0.533 (see Table 7.5). When we considered the top 31 rules that are covered by 20% of all the rules, their confidence was 0.642. Among them, 42% (13 out of 31) contained SSE information with an average confidence of 0.661. The SSE content rules were enriched among those rules exhibiting higher confidences. The same trend was also seen with unique rules: while the average confidence of 58 unique rules was 0.536, that of the 16 unique SSE rules was 0.622. Here we infer that SSE content in interaction sites is a significant feature that permits reliable classification of the interaction types.

Table 7.5: Analysis of SSE content rules over different subsets

Subset	#. of rules	Fraction(%)	$conf_1^a$	#. of SSE rules	$conf_2^b$
SSE^c	22	14.01%	0.533	-	-
$TOPK^d$	31	19.75%	0.642	13	0.661
$Unique^e$	58	36.94%	0.536	16	0.622

[a] $conf_1$: Average confidence of a rule subset;
[b] $conf_2$: Average confidence of SSE content rules in a rule subset;
[c] SSE: Association rules encoding SSE content;
[d] $TOPK$: Top K rules covering top 20% in confidence;
[e] $Unique$: Unique rules.

7.4.3 Interpretation of Discovered Association Rules

Identification of Important Rules

To select a set of informative and discriminative rules for the extraction of knowledge, most of the existing approaches rank the association rules based on the confidence value of a individual rule. A strong rule which is highly confident and represents general knowledge, may not be a good discriminative rule for the classification. Instead, a better measure of the importance of a rule should

include the following factors considered together: correlation between a property and a class, the degree of classification power, confidence and support, top K coverage and uniqueness of a rule. As noted in the previous section, the inclusion of the SSE content information in our ARBC approach has a positive effect on the classification accuracy (Table 7.4). The importance of a rule can be quantified by integrating the various factors including the SSE content information. We defined an importance factor "I" in Tables 7.6 and 7.7, by an average value of all the factors. In order to illustrate the informativeness of the rules in understanding interface features, some representative rules within the top 30% (ranked higher than 48) of factor "I" are listed in Table 7.6. The list was complemented by some other rules ranked below 48 in order to explain overlapping rules and compare association rules to rules generated from a decision tree. Similarly, rules describing the ENZ type with varying different structural features are listed in Table 7.7. Rules in Tables 7.6 and 7.7 are sorted by Type and factor "I".

We have shown that the interaction sites were dominated by non-regular region: especially for ENZ interactions, almost $\frac{2}{3}$ of the sites in average were composed of non-helix and non-beta strand regions (Figure 7.1). This is manifested in rules 29 (Table 7.7), 1, 4 and 6, all of which require $50-80\%$ content of non-regular regions to be classified as ENZ. Some of the rules containing negation predicates are strong indicators of certain interaction types. For example, "*Nohelix*" and "*Nostrand*" in the interaction sites imply ENZ (Rule 29) and nonENZ (Rules 7, 12 and 15), respectively. HET is characterized by relatively small portions of strands (Rules 18, and 19) and "*Nostrand*" (Rule 24). It is also observed that rules containing such SSE content information conjuncted with other properties (Rules 29, 7, 12, 15 and 24 in Figure 7.2) or combined with other rules (Figure 7.3(a), (b) and (c)) become stronger discriminators for

classifying PPI types than rules containing only SSE content information (Rules 1, 2, 4, 6, 14, 18, 19 and 21 in Figure 7.2). We note that some rules (Rules 29 and 7 in Figure 7.2) containing SSE information with SCOP classes are the most discriminative and informative in order to characterize ENZ and nonENZ.

Table 7.6: Representative examples of association rules for each PPI Type

#[a]	O[b]	Rule description[c]	Type[d]	Conf[e]	Supp[f]	C[g]	G[h]	K[i]	U[j]	S[k]	I[l]
1	3	If 77.31 ≤ Loop < 80.56	ENZ	0.811	0.032	1	0.214	1	1	1	0.722
2	8	If 17.57 ≤ Helix < 20.87	ENZ	0.545	0.032	1	0.102	1	1	1	0.668
3	9	If SCOPClass = 7	ENZ	0.725	0.053	1	0.184	1	1	–	0.660
4	26	If 67.59 ≤ Loop < 70.83	ENZ	0.526	0.032	–	0.048	1	1	1	0.601
5	28	If 461.83 ≤ df-ASA < 681.42 AND 2.3 ≤ LCS < 2.73	ENZ	0.625	0.032	–	0.120	1	1	–	0.555
6	37	If 57.87 ≤ Loop < 61.11	ENZ	0.467	0.037	–	0.045	–	1	1	0.510
7	2	If SCOPClass = 1 AND 12.25 ≤ nFrag < 16 AND NoStrand	nonENZ	0.882	0.032	1	0.250	1	1	1	0.738
8	11	If .66 ≤ inPro < .87	nonENZ	0.597	0.042	1	0.129	1	1	–	0.628
9	15	If 26.74 ≤ nAA < 35.32 AND 901.01 ≤ df-ASA < 1120.6	nonENZ	0.556	0.032	1	0.133	1	1	–	0.620
10	18	If SCOPClass = 1 AND 1.87 <= LCS < 2.3 9	nonENZ	0.545	0.032	1	0.137	1	1	–	0.619
11	20	If 1.43 ≤ LCS < 1.87	nonENZ	0.556	0.042	1	0.074	1	1	–	0.612
12	21	If NoStrand AND 1.87 ≤ LCS < 2.3	nonENZ	0.515	0.037	–	0.113	1	1	1	0.611
13	36	If 58.11 ≤ ASAPR < 59.52	nonENZ	0.476	0.032	1	0.065	–	1	–	0.515
14	38	If 41.67 ≤ Loop < 44.91	nonENZ	0.423	0.032	–	0.046	–	1	1	0.500
15	40	If SCOPClass = 1 AND NoStrand	nonENZ	0.484	0.064	–	0.074	–		1	0.406
16	46	If 125.14 ≤ nAtom < 165.52 AND 901.01 ≤ df-ASA < 1120.6	nonENZ	0.412	0.037	–	0.050	–	1	–	0.375
17	64	If .42 ≤ HH < .44	nonENZ	0.347	0.037	–	0.009	–	1	–	0.348
18	5	If 7.78 ≤ Strand < 10.27	HET	0.660	0.037	1	0.141	1	1	1	0.691
19	7	If 2.8 ≤ Strand < 5.29	HET	0.565	0.037	1	0.089	1	1	1	0.670
20	12	If 205.9 ≤ nAtom < 246.28	HET	0.574	0.037	1	0.143	1	1	–	0.626
21	25	If 44.91 ≤ Loop < 48.15	HET	0.479	0.037	1	0.110	–	1	1	0.604
22	32	If 3.6 ≤ LCS < 4.03	HET	0.461	0.037	1	0.100	–	1	–	0.520
23	33	If .44 ≤ HH < .46	HET	0.467	0.045	1	0.070	–	1	–	0.516
24	63	If SCOPClass = 1 AND NoStrand	HET	0.282	0.037	–	0.074	–	–	1	0.348
25	31	If SCOPClass = 3 AND 2.3 ≤ LCS < 2.73	HOM	0.470	0.033	1	0.100	–	1	–	0.521
26	98	If 3.17 ≤ LCS < 3.6	HOM	0.337	0.035	–	0.034	–	–	–	0.135
27	133	If 26.74 ≤ nAA < 35.32	HOM	0.237	0.039	–	0.041	–	–	–	0.106

Representative examples of 27 rules within top 30% are listed by sorting Columns Type and I. Rules of which order is below 48 are added for explaining overlapping rules and the comparison to rules produced from a decision tree.
[a]♯: Rule identifier;
[b]O: Order of a rule ranking by importance factor;
[c]Rule description: The body of a rule;
[d]*Type*: The head of a rule representing a PPI type;
[e]$Conf$: Confidence of a rule;
[f]$Supp$: Support of a rule;
[g]C: Rules selected from correlation-based feature subset selection (Hall 1998);
[h]G: The worth of a rule by measuring the gain ratio (Quinlan 1993) with respect to PPI types;
[i]K: Top K rules ranked within top 30%;
[j]U: Unique rules;
[k]S: SSE content rules;
[l]I: Importance factor of a rule calculated by an average of all factors such as Conf, Supp, C, G, K, U and S; "–" is replaced with value 0 when the importance factor was calculated.

Table 7.7: Representative examples of ENZ Type, presenting different structural features

#	O	Rule description	Subtype	Conf	Supp	C	G	K	U	S	I
28	24	If NoHelix	ENZ_A, ENZ_B, ENZ_C	0.508	0.069	–	0.058	1	1	1	0.606
29	1	If SCOPClass = 7 AND NoHelix	ENZ_A, ENZ_B	1.000	0.032	1	0.315	1	1	1	0.764
30	17	If 461.83 ≤ df-ASA < 681.42 AND NoHelix	ENZ_A, ENZ_B	0.593	0.037	–	0.085	1	1	1	0.619
31	39	If 461.83 ≤ df-ASA < 681.42	ENZ_A, ENZ_B	0.477	0.111	1	0.076	–	–	–	0.416
32	16	If NoHelix AND nFrag < 4.75	ENZ_A	0.612	0.032	–	0.076	1	1	1	0.620
33	19	If 4.75 ≤ nSSE < 6.62 AND NoHelix	ENZ_A	0.588	0.032	–	0.072	1	1	1	0.538
34	51	If 461.83 ≤ df-ASA < 681.42 AND 4.75 ≤ nSSE < 6.62	ENZ_A	0.417	0.032	–	0.018	–	1	–	0.367
35	77	If 44.38 ≤ nAtom < 84.76 AND 461.83 ≤ df-ASA < 681.42	ENZ_A	0.396	0.058	–	0.023	–	–	–	0.159
36	34	If 9.58 ≤ nAA < 18.16 AND 44.38 ≤ nAtom < 84.76 AND 461.83 ≤ df-ASA < 681.42	ENZ_A	0.500	0.032	–	0.045	1	1	–	0.515
37	60	If 18.16 ≤ nAA < 26.74 AND 44.38 ≤ nAtom < 84.76	ENZ_A	0.357	0.032	–	0.015	–	1	–	0.351
38	10	If 84.76 ≤ nAtom < 125.14 AND 461.83 ≤ df-ASA <681.42	ENZ_B	0.617	0.053	1	0.145	1	1	–	0.636
39	13	If 12.66 ≤ sRatio < 15.06 AND 461.83 ≤ df-ASA < 681.42	ENZ_B	0.600	0.032	1	0.113	1	1	–	0.624
40	14	If 461.83 ≤ df-ASA < 681.42 AND 10.38 ≤ nSSE < 12.25 AND SCOPClass = 2	ENZ_B	0.857	0.032	–	0.230	1	1	–	0.624
41	27	If SCOPClass = 2 AND 461.83 ≤ df-ASA < 681.42 AND 84.76 ≤ nAtom < 125.14	ENZ_B	0.789	0.032	–	0.176	1	1	–	0.599
42	35	If 10.38 ≤ nSSE < 12.25 AND 12.25 ≤ nFrag < 16	ENZ_B	0.500	0.032	–	0.043	1	1	–	0.515
43	73	If 84.76 ≤ nAtom < 125.14 AND SCOPClass = 2	ENZ_B	0.408	0.042	–	0.043	–	–	–	0.164
44	114	If 84.76 ≤ nAtom < 125.14 AND 26.74 ≤ nAA < 35.32	ENZ_B	0.307	0.037	–	0.024	–	–	–	0.123
45	109	If 681.42 ≤ df-ASA < 901.01	ENZ_C	0.317	0.048	–	0.013	–	–	–	0.126
46	137	If 84.76 ≤ nAtom < 125.14 AND 681.42 ≤ df-ASA < 901.01	ENZ_C	0.252	0.032	–	0.009	–	–	–	0.098
47	146	If SCOPClass = 4	ENZ_C	0.221	0.042	–	0.011	–	–	–	0.091
48	101	If 35.32 ≤ nAA < 43.9 AND 125.14 ≤ nAtom < 165.52	ENZ_D	0.323	0.032	–	0.041	–	–	–	0.132
49	130	If SCOPClass = 3	ENZ_D	0.238	0.069	–	0.016	–	–	–	0.108
50	141	If 901.01 ≤ df-ASA < 1120.6	ENZ_D	0.207	0.032	–	0.050	–	–	–	0.096
51	54	If 1120.6 ≤ df-ASA < 1340.19	ENZ_E	0.392	0.042	–	0.018	–	1	–	0.363

abbreviation of column names is the same as that of Table 7.6.

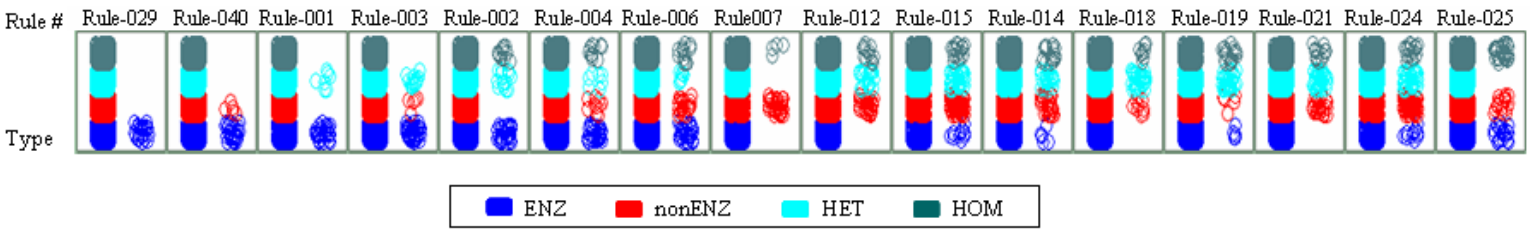


Figure 7.2: A scatter Plot matrix for PPI types and association rules. This scatter plot matrix shows clusters as collection of points separated by association rules encoding SSE content information or a SCOP class. Different colors of the left in each plot (a cell) correspond to four PPI types. The right of a plot area presents the distribution of points met with a rule on the head of a cell. Rules 29, 40, 1, and 3 separate ENZ and nonENZ from other types remarkably with few errors. The Rule 1 is a strong discriminator to classify ENZ from other types completely

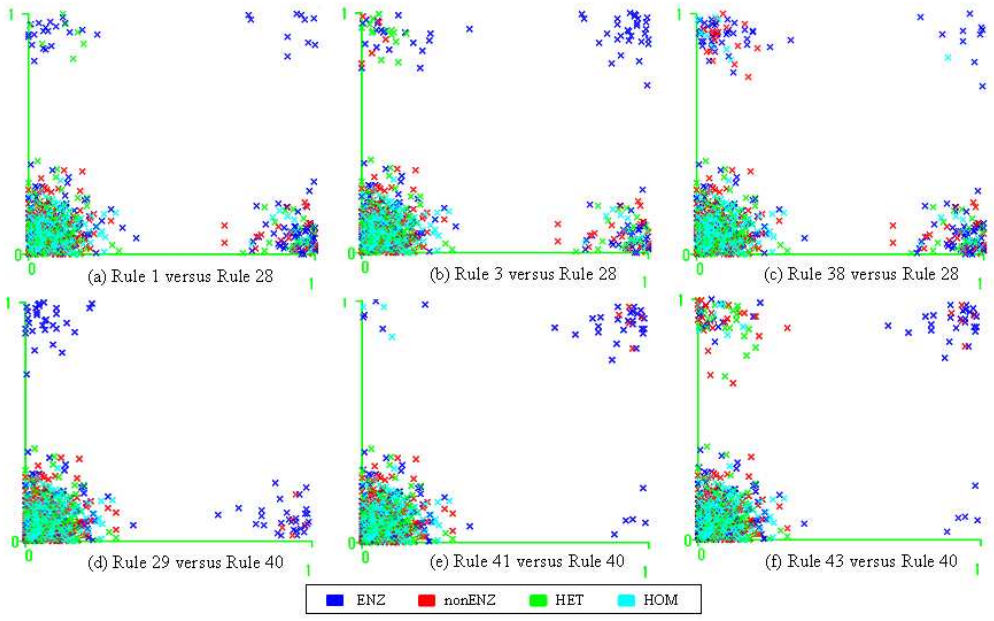


Figure 7.3: 2D plots for pairs of association rules. These plot data points by pairs of association rules. X and Y axes are a pair of rules and each of them have two boolean values. 0 represents negative data points not meeting with a rule of each axis and 1 represents for positive data points meeting with the rule. The data points on the upper left corner meet a rule used for Y axis and the data points on the down right corner meet a rule used for X axis. The points on the upper right corner meet with both rules used for X and Y axes.

Inference of Subtypes

Some rules which share the same sets of properties but differ in their value ranges or have other properties can be effective in order to compare features of different interaction types or to identify subtypes in a PPI type. For example, among the top 30% rules, Rules 38 (Table 7.7) and 16 (Table 7.6) describe types ENZ and nonENZ respectively, using the same set of properties such as number of atoms and df-ASA. However, their values imply that the interaction sites of nonENZ (Rule 16) are generally larger than those of ENZ (Rule 38). The ranges of size scales of interaction sites in ENZ are presented in Rules 35, 38 and 46

(Table 7.7) that share the same set of properties but differ in their values. The overall size of interaction sites in ENZ are described by Rule 38 with the highest confidence among those rules encoding the size of interaction sites. These are interesting cases where the structural difference between types can be directly inferred and subtypes of a PPI type can be derived by grouping different features of interaction sites. We deduced five subtypes of ENZ and a hierarchical tree (Figure 7.4) to account for those subtypes. We compiled a list of representative association rules (Table 7.7) to show structural features different among these subtypes.

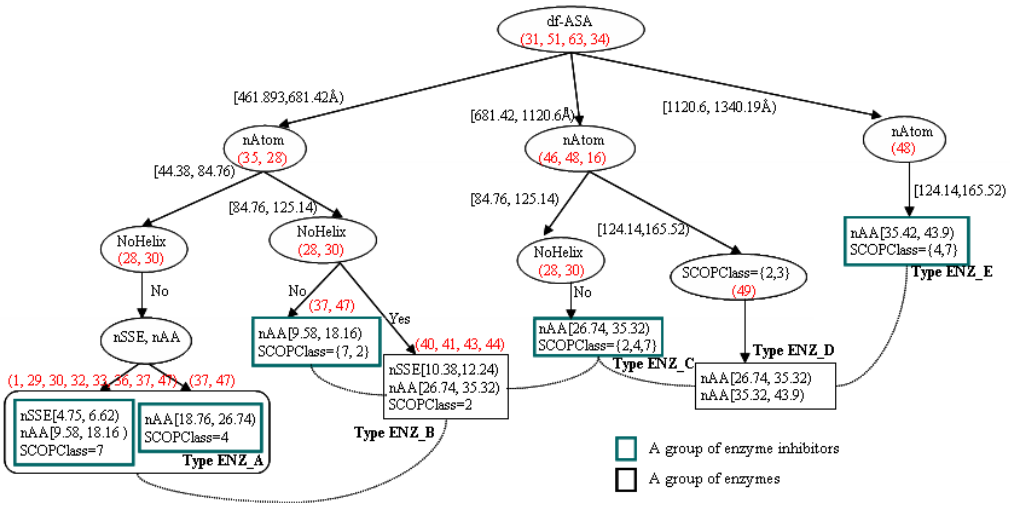


Figure 7.4: A hierarchical tree for supporting inference of subtypes. A hierarchical tree drawn from association rules represents different structural groups in ENZ. Enzyme-inhibitor interactions are characterized with size scales of interaction sites (number of atoms and df-ASA) and SSE content information (helix content). These differences of structural groups result in subtypes of PPIs. Letters in red are identifiers of rules to split branches of a tree. Dashed lines show interaction between enzymes and inhibitors in different subtypes

We note that interaction sites of enzymes are distinguished from those of enzymes-inhibitors. Interaction sites for enzyme-inhibitors are small i.e., mainly $< 1000\mathring{A}$ (Rules 34, 35, 37, 38 and 46), and are made up of strands (Rule 41)

and mostly non-regular regions (Rules 1, 4 and 6) without helix content (Rule 3, 28, 29, 30, 32, and 33) which is very informative in order to characterize enzyme-inhibitors. Remarkably Rules 30 and 28 generalize common features of inhibitors with respect to the size of interaction sites and SSE content. As Rule 29 was considered to be very discriminative to differentiate ENZ from other types, it can depict characteristics of a small group of inhibitors with indicating that enzyme-inhibitors in SCOP class 7 do not contain helix in interaction sites, (see Figure 7.3(a), (b) and (c)).

In contrast, enzymes have larger interaction sites than their inhibitors and form mixtures of helices and strands in interaction sites (Rules 40, 48, 49, 50 and 51). Both Rules 33 and 40 show that enzymes (Rule 40) have SSEs twice as many as inhibitors (Rule 33). This indicates that both enzymes and inhibitors may contain mainly strands as regular SSEs in interaction sites since enzymes are included in SCOP class 2 (mainly β) and inhibitors do not contain helices in interaction sites. This suggests that non regular regions and beta strands are mainly involved in the interfaces of enzyme-inhibitor interactions. Such extracted information can be useful for the prediction of interaction sites for enzyme-inhibitor complexes. This observation is demonstrated by some small inhibitors in Type ENZ_A (1tabi_, 2ptci_, and 4sgbi_) and Type ENZ_B (1$mcti$_). Those inhibitors interact with enzymes in Type ENZ_B. The enzymes described by Rules 40, 41 and 43 are included in SCOP superfamily trypsin-like serine proteases (2.47.1) and the inhibitors are mainly in SCOP class 7 which is composed of small proteins dominated by metal ligand, heme, and disulfide bridges.

It is possible in a similar wayto infer subtypes of other PPI types. Among PPI types, ENZ has plenty of rules (a total of 65) to derive subtypes. Hence, the comparative analysis of association rules was presented for ENZ.

Comparison of Association Rules to PART Rules

To improve our understanding of the association rules discovered, we compared PART rules produced from a decision tree built using C4.5 over our properties with the association rules. There were a total of 44 PART rules generated and their average confidence and support were 0.99 and 0.02 respectively. We have collected a representative list of PART rules in Table 7.9. In the comparison of the association rules with PART rules, PART rules are more complicated with the composition of more predicates in rule bodies than those in association rules. Typically, one PART rule corresponds to more than $2 \sim 3$ association rules (Table 7.9). Both rules provided quantitative descriptions. However, property values in PART rules represent split points for classification and are not represented by intervals of quantitative values. Some PART rules (Rules 1, 3 and 38 in Table 7.9) including identical properties with different split points in the same rule bodies were not clear enough to determine decision boundaries of properties. These limit the readability and understandability of PART rules whilst the association rules were simple enough to be interpreted by users. It was also possible with association rules to support the comparative analysis of rules between different PPI types as we inferred the possibility of subtypes and relative information by comparison of size scales of interaction sites in ENZ. A set of association rules discovered by ARM comprises mostly weak rules together with a small number of strong rules. On the contrary, most PART rules consist of a number of very strong rules which have the highest confidences and low supports.

One of the most notable differences between association rules and PART rules is in how to handle overlapping rules between different types. If two different interaction types are predicted from the identical head of a rule, these are called overlapping rules. There were 99 such cases out of a total of 157 rules (Table 7.3).

Table 8 shows representative examples of overlapping rules. Examination of the overlapping rules shared by ENZ and nonENZ indicated that these types are similar in terms of df-ASA, nAtom, and nAA (Table 8) differentiated by combination with the rest of properties such as SSE content, average length of consecutive residues, size ratio, and hydrophobicity. PART rules are unique cross PPI types.

Table 7.8: PRepresentative examples of overlapping association rules

#[a]	#[b]	Rule description[c]	Types[d]	Conf[e]	Supp[f]	Conf[g]	Supp[h]
52	43	If 84.76 $\leq$ nAtom $<$ 125.14 AND SCOPClass = 2	ENZ[1] OR nonENZ[2]	0.408	0.042	0.306	0.032
53	35	If 44.38 $\leq$ nAtom $<$ 84.76 AND 461.83 $\leq$ df-ASA $<$ 681.42	ENZ[1] OR nonENZ[2]	0.396	0.058	0.252	0.037
54	48	If 35.32 $\leq$ nAA $<$ 43.9 AND 125.14 $\leq$ nAtom $<$ 165.52	ENZ[1] OR nonENZ[2]	0.323	0.032	0.376	0.037
55	46	If 84.76 $\leq$ nAtom $<$ 125.14 AND 681.42 $\leq$ df-ASA $<$ 901.01	ENZ[1] OR nonENZ[2]	0.252	0.032	0.336	0.042
56	26	If 3.17 $\leq$ LCS $<$ 3.6	HET[1] OR HOM[2]	0.357	0.037	0.337	0.035

Examples of overlapping rule are selected from Tables 6 and 7.
[a]♯ Rule identifier;
[b]♯: Rule identifier in Tables 6 and 7;
Rule description[c]: The body of overlapping rules between the two types;
[d]$Types$: PPI $Type^1$ and $Type^2$ having overlapping rules in common;
[e,g]$Conf$: Confidences of overlapping rules for $Type^1$ and $Type^2$ respectively;
[f,h]$Supp$: Supports of overlapping rules for $Type^1$ and $Type^2$ respectively.

Table 7.9: PART rules generated by decision trees using $C4.5^a$

#[b]	Rules discovered by C4.5 Decision Tree	Type	Conf	Supp	Corresponding rules[c]
5	AVGASA > 68.73025 AND nAtom > 60 AND LCS > 2.611111 AND Strand ≤ 32.857 AND SCOPClass = 7	ENZ	1	0.03	35, 5, 3, 36
38	sRatio ≤ 29.411765 AND HH > 0.277096 AND SCOPClass = 2 AND Strand > 16.949 AND Strand > 21.324 AND nSSE > 10	ENZ	1	0.02	40, 39
4	Loop > 50.299 AND nAtom > 60 AND Helix ≤ 33.636 AND AVGASA ≤ 41.137133	ENZ	0.99	0.07	35, 6
27	inPro ≤ 2.016077 AND Helix > 48.485 AND LCS > 1.727 AND Strand ≤ 8.571 AND SCOPClass = 1 AND AVGASA ≤ 53.133	nonENZ	1	0.02	8, 10
40	SCOPClass = 1 AND Strand ≤ 2.26	nonENZ	1	0.01	15
1	nAtom > 189 AND Loop ≤ 66.316 AND nSSE > 13 AND Helix > 19.481 AND sRatio ≤ 80.833 AND inPro > -1.570 AND LCS > 3.714 AND Loop ≤ 46.7	HET	1	0.05	20, 21
3	nAtom > 212 AND Strand ≤ 10.738 AND nSSE > 13 AND inPro > -1.476973 AND nAtom > 384	HET	1	0.05	20, 18, 19
34	SCOPClass = 3 AND Helix > 18.421	HOM	1	0.02	25
15	HH > 0.433 AND AVGASA > 55.984 AND nAA ≤ 34	HOM	1	0.01	27

[a]: A total of 44 rules produced by a decision tree using C4.5 algorithm in WEKA machine learning library;
[b]♯: PART rule identifier;
[c]Corresponding rules: Association rule identifiers (Tables 7.6, 7.7 and 7.9)

7.5 Conclusions

We have developed a classification method that categorizes each PPI into one of four different types using association rule based classification (ARBC). The application of association rule mining over 354 known PPI types using 14 properties yielded a total of 157 rules, which in turn discriminated the features of interaction sites for different PPI types and were used to generate a classification model to predict PPI types. Our ARBC approach performed competitively compared with conventional methods applied directly to the property values: for example, the work in (Zhu et al. 2006) reported an accuracy of 91.8% for the classification of three types of interactions by directly applying SVM. Although it is not possible to make a direct comparison of their method with ours due to heterogeneity of the data set, this suggests that the processes of association rule generation and subsequent pruning do not incur a loss of relevant information. At the same time, our results demonstrated that we were able to considerably improve the accuracy of the prediction of PPI types through the use of structural domain information for the description of interaction interfaces, and also the use of secondary structure content. Although SSE content alone could not classify interaction sites with high accuracy, its incorporation with other properties improved the accuracy of classification.

Our approach based on ARBC has a clear advantage over conventional methods because results are reported in terms of rules that are a quantitative description of properties and hence their interpretation is straightforward and simple. Thus, biologists can easily judge if a discovered rule is interesting or not for further investigation. Analysis of common and unique properties together is a unique feature of our approach, unlike conventional classification methods which typically capture unique properties only. Common rules capture those properties which are common between PPI types. In particular enzyme inhibitor (ENZ)

and non-enzyme inhibitor (nonENZ) interactions, both being non-obligate or transient, share more properties in common than with other types. As we have demonstrated, all these features produce descriptive rules, enabling their simple and powerful interpretation. We observed that the property distributions of homo-obligate interactions are similar to those of hetero-obligate interactions but distinct from those of non-obligate interactions. We found that obligate interactions have larger and more hydrophobic interaction sites than non-obligate interactions. Hydrophobic residues including Leu, Ala, and Val were found more frequently in obligate interactions whilst polar residues including Ser and Gly were present in non-obligate interactions. Charged residues (Glu, Asp, Lys, and Arg) were seen frequently in all interaction types. On the basis of a detailed analysis of association rules, it was observed that interactions between enzymes and their inhibitors were separated into several different structural subgroups. This may lead to the possibility of different subtypes of PPIs being involved in transient interactions. Our findings based on the interpretation of association rules are consistent with the description of homo-obligate complexes in previous studies (Nooren and Thornton 2003, Zhu et al. 2006).

In future work we plan to improve our approach by incorporating additional properties such as energy functions and electric potentials for the generation of more accurate and meaningful association rules. The unique contribution of our work is the development of a novel methodology that analyzes specificities and commonalities for interaction types, and we intend to extend this to the prediction of interaction partner and interaction sites. Most of the work associated to this chapter has been included in a publication recently submitted to the journal BMC Bioinformatics in (Park et al. 2009).

7.6 Summary

In this chapter we have presented a new computational approach for the prediction of PPI types based on protein complex structure information. Our approach utilized association rule based classification (ARBC) to correctly classify types and characterize PPI binding sites.

Owing to complexity of experimental techniques, at present there is a reduced number of available protein complex structures. There are not enough examples available to develop an integrative learning approach as in previous chapters. However, in the near future, it is expected that the number of crystalized protein structures will increase. In this case, the information related to PPI types can be useful to enhance the predictions made by our previous techniques (one-class classification). Consequently, the ARBC approach described in this chapter aims not only to deal with the problem of prediction of PPI types but also to develop and initial framework to generate new useful data for the task of prediction of PPI in a broader context.

Chapter 8

Conclusions and Future Work

Protein-protein interactions (PPI) operate at every level of cellular function. The correct identification of these interactions is important to systematically understand the roles played by cellular proteins in diverse biological functions. Consequently, the prediction of protein-protein interactions (PPI) has emerged recently as an important problem in the fields of bioinformatics and systems biology.

Large scale biological experiments for identification of PPI can directly detect hundred or thousands of protein interactions at a time. However the resulting data sets are often incomplete and exhibit high false-positive and false-negative rates. On the other hand, small scale experiments for identification of new PPI are more accurate but are expensive and time consuming, and consequently it is not feasible to test every possible PPI. For all these reasons, there has been an increasing need to develop computational approaches, especially in the machine learning investigation area, to improve our knowledge about this type of biological interactions. In general proteins do not work alone but in groups called protein complexes. In this thesis we focussed specifically in the prediction of co-complex interactions, where the objective is to identify and characterize protein pairs which are members of the same protein complex.

Several studies have been developed in the past based on the integrative learning analysis of diverse biological sources of information. Several machine learning techniques, mostly supervised learning approaches, have been employed to improve the accuracy and trustability of predicted protein interacting pairs. They have demonstrated that the combined use of direct and indirect biological insights can improve the quality of predictive PPI models. The prediction of PPI has been commonly viewed as a binary classification problem (whether any two proteins do or do not interact). However the nature of the data creates two major problems which can affect results. These are firstly imbalanced class problems due to the number of positive examples (pairs of proteins which really interact) being much smaller than the number of negative ones. Secondly the selection of negative examples can be based on some unreliable assumptions which could introduce some bias in the classification results.

In the first part of this thesis (Chapter 4) we addressed these common drawbacks by exploring the use of one-class classification (OCC) methods to deal with the task of prediction of PPI. According to our knowledge, when we initiate this research, OCC models have never been employed before to deal with this predictive task. OCC methods utilize examples of just one class to generate a predictive model which consequently is independent of the kind of negative examples selected; additionally these approaches are known to cope with imbalanced class problems. We designed and carried out a performance evaluation study of several OCC methods for this task. Among available techniques employed, we found that the Parzen density estimation approach clearly exhibited the best performance. We then undertook a rigorously comparative performance evaluation between the Parzen OCC method and several conventional learning techniques, which had been employed before to deal with the task of PPI predic-

tion. For this evaluation, different scenarios were considered, for instance varying the number of negative examples used for training purposes. We demonstrated that the Parzen OCC approach performs very competitively and outperforms the rest of conventional classifiers in most of the situations, up to the case where the ratio of negative to positive examples is approximately 25 to 1.

The performance of conventional binary classification approaches is highly influenced by the quantity of negative examples used to train the respective models. Thus, classification models generated from these type of techniques are more reliant on negative information (in this case an untrustworthy set of negative PPI examples) than on positive information (experimentally corroborated PPI examples). Our results indicate that the problem of prediction of PPI can indeed be formulated as an OCC problem where the predictive model is based on real (trustworthy) PPI data. In the specific case of prediction of co-complexed proteins, we found that the Parzen OCC method is able to generate models which perform competitively with those generated by conventional classifiers, independently of the quality and quantity of the negative examples available. Most of the work associated with this chapter has been included in a referred publication in (Reyes and Gilbert 2007).

Further in our investigation we addressed a new drawback which appears to be affecting the performance of the PPI prediction task (Chapter 5). This is associated with the composition of positive gold standard set (set of protein pairs that really interact), which contain a high proportion of examples (2/3 of the total) related to ribosomal protein. The ribosomal-based protein complexes contain a large number of proteins. We demonstrated that this situation indeed biases the classification task, resulting in an over-optimistic performance results. The models generated employing all examples are biased toward the prediction

of ribosomal-based PPI. It was observed that the problem of prediction of non-ribosomal PPI is a much difficult task.

The problem associated with the high proportion of ribosomal-based proteins has not been previously reported or addressed according to the best of our knowledge. Consequently, we focused our research to deal with this specific subtask, aiming to improve the performance associated to the prediction of non-ribosomal PPI when using the Parzen OCC model. We investigated the effect of integrating new biological information into the classification process, based on data from mRNA expression experiments and protein secondary structure (SS) information. We demonstrated that the integration of data from diverse mRNA expression experiments into a single data set has a negative effect on the performance of the Parzen OCC approach. There is no synergy effect in this case, and Parzen OCC models based on individual mRNA expression experiment outperform the one which integrates all the data. The integration of protein secondary structure information significantly improves the performance of the Parzenn OCC approach for this predictive task. The performance of all of the models evaluated is improved when SS-based features are incorporated into the classification process, including the case when no mRNA expression data is used.

Based on previous results obtained, we investigated several strategies to combine predictions of different Parzen OCC models induced from diverse subsets of biological data. The hypothesis behind this approach, was related to the observations that single mRNA and SS-based information seems to incorporate independent insights to the PPI prediction problem. Four models were selected for this procedure, three based on individual mRNA expression experiments (without SS information) and one based on SS information (without mRNA expression data). We demonstrated that these models exhibited a high degree of diversity in their predictions, corroborating our assumption. We also

demonstrated that it is possible to significantly improve the performance of the prediction of non-ribosomal PPI by combining the predictions of several Parzen OCC models. The weighted average combination approach exhibited the best performance, and also gave some insights regarding the relative importance of the different classification models employed. Most of the work associated to this chapter has been included in a referred publication in (Reyes and Gilbert 2008).

Further in this thesis we undertook a preliminary evaluation analysis of the capability of the Parzen OCC approach to predict new potential PPI targets (Chapter 6). The final goal associated with the use of computational methods for predicting PPI is to predict or identify new potential PPI targets. These potential targets can then be used, for instance, to guide biologists developing small scale experiments in order to validate them.

Employing the classification approach we previously generated based in the combination of several and diverse Parzen OCC models, we were able to predict a set of 818 new PPI for further analysis. With this validation study we intend to look for evidence that support the new PPI predictions. Firstly, we searched in the literature and related databases for experimental evidence related to these new predictions. We found that many of them are supported by experiments associated to the identification of PPI complex. Secondly, we analyze the topological properties of the PPI network associated to these new PPI predictions. Computationally, a protein-protein interaction network could be conveniently modeled as an undirected graph, where the nodes are proteins and two nodes are connected by an undirected edge corresponding to a certain kind of of interaction. We found that our predicted PPI network, share several common properties that have been recently associated to many biological networks, especially PPI networks. Finally, we focused in the identification of

highly connected groups of proteins within our predicted PPI network. These groups or clusters can be associated for instance to novel protein complexes. We were able to identify, describe and analyze three important clusters. The initial analysis of these clusters, showed certain biological evidence supporting the PPI predictions generated using the Parzen OCC approach.

Finally in this thesis we explored a slightly different area related with the prediction of PPI types (Chapter 7). This is associated with the classification of PPI structures (complexes) contained in the Protein Data Bank (PDB) data base according to its function and binding affinity.

For this we implemented a computational approach for the prediction of PPI types employing association rule based classification (ARBC), which includes association rule generation and posterior classification based on the discovered rules. Our approach based on ARBC has a clear advantage over conventional methods because results are reported in terms of rules that are a quantitative description of properties and hence their interpretation is straightforward and simple. Thus, biologists can easily judge if a discovered rule is interesting or not. The unique contribution of our work is the development of a novel methodology that analyzes specificities and commonalities for interaction types. The analysis of common and unique properties together is a unique feature of our approach, unlike conventional classification methods which typically capture unique properties only. Common rules capture those properties which are common between PPI types. Due the relatively reduced number of crystalized protein complexes available, it is not possible at the moment to link the results and biological features of this task with the one studied before related with the prediction of PPI complexes. However this could be possible in the near future when more PPI structures will be available. The results associated to this investigation has been

included in a publication recently submitted to the journal BMC Bioinformatics (Park et al. 2009).

Computational learning of protein-protein interactions and protein interaction networks is still an undergoing research topic. Many important questions and drawbacks related with these tasks remain as open challenges. In this sense we can visualize some future directions to guide our work, as follows:

- In the future we will extend the use of OCC models for prediction of PPI complexes in other species, for instance in human-based protein complexes.
- We will investigate the incorporation of other kind of biological information, in order to improve the performance of classifiers for PPI prediction. Specifically, we are interested in the use of protein structural information.
- We are also interested in the evaluation of novel strategies for the combination of diverse classification models. An example of this is the "stacked generalization" approach, which according to our knowledge, has never been utilized for the combination of predictions made by OCC models.
- An important area for future work, is to continue with the validation of new PPI predicted in this thesis. We are specifically interested to develop an in-depth analysis of the protein clusters discovered in the predicted PPI network. In order to gain biological knowledge about these inferred relationships.
- We intend to improve our approach based on ARBC for prediction of PPI Types, by incorporating additional properties such as energy functions and electric potentials, aiming to generate more accurate and meaningful association rules.

Appendix A

List of predicted PPI for further validation

The following table exhibit the complete list of new 818 PPI predicted in our research, as showed in Chapter 6. The classification model utilized is based in the combination of several parzen OCC classifiers, generated from diverse sets of biological data:

No	ID-1	ID-2	No	ID-1	ID-2
1	YDL126C	YDR394W	11	YER021W	YKL210W
2	YDL126C	YDR427W	12	YER021W	YOR249C
3	YDL126C	YER021W	13	YDL126C	YOR362C
4	YDL126C	YMR314W	14	YBL041W	YDL126C
5	YDL126C	YPR108W	15	YDL132W	YDR394W
6	YDL126C	YOR259C	16	YDL132W	YER021W
7	YDL126C	YOR157C	17	YDR177W	YER021W
8	YDL126C	YOR261C	18	YDL132W	YPR108W
9	YDL132W	YDL147W	19	YDR177W	YDR394W
10	YDR177W	YOR261C	20	YDL126C	YGL011C

No	ID-1	ID-2	No	ID-1	ID-2
21	YBL041W	YDL132W	56	YDR177W	YPR108W
22	YDL126C	YOR117W	57	YER094C	YJR099W
23	YDL126C	YFR050C	58	YJR099W	YOR261C
24	YKL210W	YOR117W	59	YDL132W	YML092C
25	YDL126C	YDL147W	60	YOR124C	YOR249C
26	YDL097C	YDL126C	61	YDR394W	YOR249C
27	YFR050C	YOR249C	62	YDL132W	YFR050C
28	YDL126C	YKL210W	63	YDL132W	YKL213C
29	YKL210W	YPR108W	64	YGR135W	YOR249C
30	YDL126C	YFR004W	65	YFR004W	YKL210W
31	YDR394W	YKL210W	66	YGL048C	YOR124C
32	YDL007W	YDL126C	67	YKL022C	YOR117W
33	YOR157C	YOR249C	68	YDL150W	YHR143WA
34	YOR249C	YOR261C	69	YDR092W	YGR253C
35	YDL132W	YOR261C	70	YGR232W	YKL210W
36	YDR394W	YLL039C	71	YDR092W	YJL001W
37	YDL132W	YDR427W	72	YDL147W	YDR177W
38	YDL132W	YGL011C	73	YDR177W	YOR259C
39	YDL007W	YOR124C	74	YKL210W	YKL213C
40	YDR427W	YOR249C	75	YJR099W	YOR117W
41	YDR394W	YOR124C	76	YDL150W	YNR003C
42	YFR004W	YJR099W	77	YJL001W	YKL213C
43	YBR082C	YDL007W	78	YKL145W	YOR124C
44	YGL011C	YOR249C	79	YDL165W	YGR005C
45	YDL126C	YOL038W	80	YKL210W	YOR362C
46	YDL132W	YOR259C	81	YDL147W	YDR092W
47	YDR092W	YOR259C	82	YBL041W	YKL213C
48	YDR054C	YOR117W	83	YHR143WA	YOR174W
49	YBR082C	YER094C	84	YDL007W	YDR092W
50	YBR082C	YOR261C	85	YDL150W	YOR224C
51	YDL126C	YML092C	86	YML111W	YPR108W
52	YKL210W	YOR157C	87	YCR093W	YHR143WA
53	YDL097C	YKL210W	88	YDL150W	YKL144C
54	YDL132W	YKL010C	89	YJL197W	YPR108W
55	YHR200W	YKL210W	90	YKL058W	YOR174W

No	ID-1	ID-2	No	ID-1	ID-2
91	YDR177W	YOR117W	126	YBR082C	YDR394W
92	YIL075C	YKL010C	127	YBR082C	YDL147W
93	YDR429C	YER025W	128	YFR052W	YJL197W
94	YIL156W	YJL001W	129	YDL132W	YOR124C
95	YDL150W	YOR210W	130	YER021W	YJR099W
96	YDR092W	YER012W	131	YDR311W	YOR174W
97	YDL126C	YKL010C	132	YDR092W	YOR261C
98	YDL097C	YJL197W	133	YGL240W	YOR261C
99	YKR094C	YPR108W	134	YDR092W	YDR394W
100	YBR058C	YDL126C	135	YDR092W	YMR314W
101	YKR094C	YOR117W	136	YJR007W	YMR309C
102	YDL007W	YDR390C	137	YGL025C	YPR056W
103	YBR049C	YOR194C	138	YER021W	YOR124C
104	YKL213C	YPR108W	139	YOR117W	YOR124C
105	YDR054C	YOR261C	140	YDR092W	YER021W
106	YDL132W	YFR004W	141	YKL144C	YPR186C
107	YFR050C	YKL213C	142	YNL062C	YOL139C
108	YBR154C	YDL150W	143	YDL147W	YIL156W
109	YDR092W	YOL038W	144	YBR049C	YDR404C
110	YBR193C	YIL143C	145	YKL210W	YPR103W
111	YKL213C	YPR103W	146	YGL025C	YOR194C
112	YER133W	YNL084C	147	YBR193C	YCR093W
113	YFR004W	YHR166C	148	YDR092W	YGL048C
114	YDR328C	YER021W	149	YDL147W	YLL039C
115	YDR394W	YJR099W	150	YDR092W	YPR103W
116	YDL147W	YOR124C	151	YHR058C	YKR062W
117	YDL150W	YNL151C	152	YER025W	YLR291C
118	YDR092W	YOR362C	153	YGR083C	YNL062C
119	YBL041W	YKL210W	154	YDL164C	YHR118C
120	YER021W	YGL240W	155	YBR082C	YFR004W
121	YIL075C	YOR249C	156	YDL147W	YJR099W
122	YER148W	YFL031W	157	YBL014C	YKL125W
123	YOR048C	YOR098C	158	YLR291C	YPL237W
124	YDL150W	YPR190C	159	YBR082C	YER021W
125	YIL046W	YOR339C	160	YER021W	YLL039C

No	ID-1	ID-2	No	ID-1	ID-2
161	YBR154C	YML112W	196	YDL132W	YGR135W
162	YKR062W	YOL051W	197	YDL132W	YOR362C
163	YKL145W	YKR094C	198	YIL156W	YOL038W
164	YLR131C	YOR224C	199	YDL097C	YKL213C
165	YDL147W	YMR275C	200	YOR151C	YPL042C
166	YDL150W	YPR110C	201	YER022W	YIL021W
167	YHR143WA	YPR186C	202	YGL025C	YKL028W
168	YBR038W	YDR182W	203	YFL031W	YPR086W
169	YDR092W	YFR052W	204	YDL150W	YOR207C
170	YGL011C	YMR275C	205	YDL007W	YLL039C
171	YDL165W	YKL028W	206	YBR154C	YPR186C
172	YLR291C	YMR309C	207	YER151C	YOR117W
173	YBR082C	YOR117W	208	YKR062W	YLR131C
174	YDR427W	YGL240W	209	YJL148W	YML043C
175	YBR079C	YDR211W	210	YGL070C	YOL135C
176	YGL048C	YKL210W	211	YML112W	YOR194C
177	YER012W	YMR275C	212	YPR186C	YPR187W
178	YDL097C	YOR124C	213	YGR186W	YNL025C
179	YER021W	YKL010C	214	YGR083C	YJR007W
180	YGL240W	YOR259C	215	YDL150W	YNL113W
181	YBL008W	YGR005C	216	YML098W	YOL135C
182	YDR308C	YJL127C	217	YER021W	YKL213C
183	YJL197W	YOL038W	218	YOR210W	YPL042C
184	YMR146C	YOL139C	219	YNL113W	YPR186C
185	YER144C	YKL022C	220	YGL025C	YML098W
186	YJL197W	YML092C	221	YBR154C	YCR093W
187	YDL097C	YJR099W	222	YNL025C	YPR168W
188	YJL056C	YOR210W	223	YCR093W	YGL070C
189	YER025W	YMR260C	224	YDL005C	YIL021W
190	YKL213C	YOR249C	225	YOR363C	YPR086W
191	YDR054C	YHR200W	226	YJR007W	YNL062C
192	YOR249C	YOR259C	227	YHR027C	YKL210W
193	YBR193C	YJL127C	228	YCR039C	YKL028W
194	YGR253C	YMR275C	229	YFL031W	YGR005C
195	YGL025C	YPR187W	230	YOR249C	YPR108W

No	ID-1	ID-2	No	ID-1	ID-2
231	YDR394W	YJL197W	266	YDR211W	YPL237W
232	YDR167W	YHR143WA	267	YKL028W	YPL042C
233	YBR097W	YDL217C	268	YBR193C	YCL066W
234	YCL066W	YPR086W	269	YNL261W	YOR217W
235	YJL001W	YJL197W	270	YBR198C	YPR187W
236	YML043C	YPR010C	271	YDR054C	YER144C
237	YGL240W	YOR157C	272	YDR429C	YPL237W
238	YER098W	YIL046W	273	YBR060C	YJL173C
239	YDR177W	YKL213C	274	YKL213C	YML092C
240	YDR167W	YGL025C	275	YDR328C	YOR117W
241	YMR275C	YOR157C	276	YGR186W	YML112W
242	YDR390C	YHR200W	277	YBR193C	YML112W
243	YJL006C	YOR151C	278	YNR003C	YPR186C
244	YCR040W	YKL058W	279	YHR143WA	YKL028W
245	YJL006C	YOR194C	280	YDR308C	YPL042C
246	YDR390C	YPR108W	281	YIL156W	YMR314W
247	YER171W	YPR168W	282	YOL005C	YOR174W
248	YJL140W	YLR071C	283	YBR057C	YDR182W
249	YCR039C	YOR174W	284	YDR328C	YHR200W
250	YDR427W	YML111W	285	YDR207C	YPL122C
251	YCR039C	YGR186W	286	YDL147W	YER100W
252	YCL066W	YDL140C	287	YGR186W	YIR017C
253	YDR092W	YGL011C	288	YFL031W	YKL028W
254	YKL058W	YLR131C	289	YER133W	YNL233W
255	YER025W	YMR146C	290	YGL025C	YPR086W
256	YBR193C	YDR146C	291	YFL031W	YOR151C
257	YML043C	YNL248C	292	YNL236W	YOL005C
258	YER094C	YLL039C	293	YMR227C	YPR168W
259	YDR054C	YDR394W	294	YKL139W	YPL046C
260	YHR058C	YOR210W	295	YDR394W	YKR094C
261	YBR193C	YJL140W	296	YKL145W	YMR275C
262	YCR040W	YOR174W	297	YHR143WA	YML112W
263	YDR167W	YIL021W	298	YDR404C	YKL058W
264	YGL048C	YJR099W	299	YDL005C	YPR187W
265	YJL140W	YPR168W	300	YDL097C	YDR092W

No	ID-1	ID-2	No	ID-1	ID-2
301	YDR308C	YHR084W	336	YGL048C	YIL148W
302	YDL126C	YGR135W	337	YBL093C	YKL058W
303	YDR211W	YJR007W	338	YGR186W	YJL056C
304	YEL032W	YOL006C	339	YBR278W	YMR001C
305	YDL132W	YER151C	340	YDL097C	YGR184C
306	YDL097C	YDR059C	341	YLL004W	YNL290W
307	YIL128W	YOR151C	342	YDL005C	YKR062W
308	YCR039C	YER022W	343	YOL115W	YPR162C
309	YBL093C	YPR187W	344	YER021W	YKR094C
310	YBL093C	YOL005C	345	YDL217C	YDR142C
311	YFR004W	YMR275C	346	YDR311W	YGL070C
312	YNL216W	YNL221C	347	YJR068W	YPR175W
313	YDR207C	YIL143C	348	YDL147W	YGR184C
314	YFL031W	YOR174W	349	YJL127C	YPR187W
315	YDL097C	YLL039C	350	YGL070C	YGR005C
316	YDR394W	YHR166C	351	YJR007W	YLR291C
317	YDL007W	YKL213C	352	YCR039C	YKL058W
318	YBL093C	YHR143WA	353	YBL084C	YDL147W
319	YBR193C	YOR363C	354	YOL005C	YPL042C
320	YLL004W	YNL312W	355	YKL010C	YOR249C
321	YNL062C	YOR260W	356	YDR429C	YMR260C
322	YJL194W	YNL312W	357	YHR166C	YOR261C
323	YBR082C	YDL097C	358	YDR211W	YNL062C
324	YGL043W	YOR174W	359	YDR394W	YIL156W
325	YER094C	YIL148W	360	YOL135C	YPL122C
326	YLR071C	YPR187W	361	YGL035C	YKR062W
327	YDL140C	YLR131C	362	YDR092W	YPR108W
328	YDR059C	YER021W	363	YKR094C	YOR261C
329	YBL041W	YMR275C	364	YDL007W	YKL210W
330	YFR050C	YMR275C	365	YHR058C	YJL140W
331	YAR007C	YJL194W	366	YGR186W	YLR131C
332	YGL025C	YKL058W	367	YIL148W	YKL145W
333	YNL025C	YOL135C	368	YBR154C	YMR270C
334	YLR131C	YPR086W	369	YGL201C	YOL094C
335	YFR004W	YKL213C	370	YKL210W	YOL038W

No	ID-1	ID-2	No	ID-1	ID-2
371	YDL140C	YJL056C	406	YDL126C	YHR200W
372	YDL064W	YER021W	407	YDL005C	YMR236W
373	YDR308C	YJL056C	408	YBR193C	YIL128W
374	YBR202W	YJR068W	409	YJL127C	YPR086W
375	YJL025W	YPR187W	410	YMR146C	YPL237W
376	YDL102W	YMR001C	411	YDR443C	YKR062W
377	YAR007C	YIL150C	412	YBR058C	YFR052W
378	YDR404C	YDR443C	413	YHR041C	YPR056W
379	YDR059C	YOR261C	414	YER012W	YJL197W
380	YJL001W	YOR249C	415	YOL005C	YOL051W
381	YGL011C	YKL213C	416	YER171W	YHR143WA
382	YDL126C	YJL001W	417	YBR058C	YLR127C
383	YKL028W	YPR187W	418	YDL005C	YPR086W
384	YKL010C	YPR108W	419	YCR040W	YER148W
385	YIR017C	YKR062W	420	YER148W	YJR063W
386	YDR182W	YNL233W	421	YJL006C	YPR086W
387	YKL058W	YOL135C	422	YJL148W	YMR270C
388	YHR084W	YKR062W	423	YDL147W	YKR094C
389	YBL008W	YKL028W	424	YER021W	YJL197W
390	YDR443C	YHR058C	425	YBL023C	YOL006C
391	YDL132W	YOL038W	426	YCR081W	YHR058C
392	YDR311W	YGR044C	427	YHR058C	YPL122C
393	YHR164C	YML065W	428	YDL017W	YOL094C
394	YHR084W	YKL058W	429	YBL008W	YPR086W
395	YGR005C	YOR038C	430	YDR404C	YOL051W
396	YJR006W	YOR217W	431	YDR087C	YJR063W
397	YOR260W	YOR361C	432	YGL035C	YKL058W
398	YER022W	YLR131C	433	YBL093C	YMR236W
399	YER025W	YOL139C	434	YDR146C	YDR308C
400	YBL084C	YER094C	435	YNL062C	YPL237W
401	YKL028W	YPR168W	436	YBR087W	YML065W
402	YBR038W	YBR109C	437	YBL041W	YOR249C
403	YDR052C	YHR164C	438	YER133W	YIR006C
404	YDL097C	YIL156W	439	YGR083C	YMR146C
405	YDR311W	YNL236W	440	YBR198C	YOL005C

No	ID-1	ID-2	No	ID-1	ID-2
441	YKL010C	YOR261C	476	YHR058C	YPL042C
442	YDR118W	YER017C	477	YBR079C	YOR260W
443	YDR404C	YJL006C	478	YGL240W	YOR117W
444	YDR443C	YOR194C	479	YGL112C	YOR224C
445	YGL025C	YGR005C	480	YML111W	YOL038W
446	YIR017C	YOR174W	481	YDL108W	YHR041C
447	YDL147W	YJL197W	482	YGR186W	YPL042C
448	YCL066W	YOR194C	483	YBR038W	YKR048C
449	YDR443C	YGL025C	484	YDL005C	YGL112C
450	YBR198C	YOR151C	485	YER144C	YIL075C
451	YDL064W	YOR261C	486	YDL008W	YER017C
452	YCL066W	YKL058W	487	YBL093C	YDL140C
453	YBR193C	YDL140C	488	YGR005C	YIR017C
454	YLR291C	YNL062C	489	YGR083C	YMR309C
455	YER025W	YGR083C	490	YBR193C	YGR005C
456	YCR040W	YKR062W	491	YER148W	YIL021W
457	YER022W	YGL035C	492	YIL148W	YOR117W
458	YDL140C	YGL043W	493	YDL132W	YMR314W
459	YFR004W	YKR094C	494	YBR058C	YPR103W
460	YBR193C	YGL043W	495	YBR198C	YOR210W
461	YJL127C	YOR174W	496	YDL147W	YIL148W
462	YHR041C	YPR086W	497	YBL093C	YGR186W
463	YDR177W	YFR004W	498	YER148W	YOR038C
464	YER094C	YKR094C	499	YIL148W	YOR261C
465	YGL011C	YJL197W	500	YBR193C	YPR025C
466	YGL153W	YNL131W	501	YDR328C	YPR108W
467	YCL066W	YKR062W	502	YCL066W	YER022W
468	YJR007W	YMR146C	503	YER148W	YJL056C
469	YOR210W	YPR186C	504	YER148W	YKL144C
470	YBL008W	YKR062W	505	YDR146C	YOR174W
471	YJL056C	YKL058W	506	YBL023C	YNL088W
472	YJR007W	YOR260W	507	YLR274W	YNL088W
473	YER022W	YPL042C	508	YJL194W	YOL094C
474	YBL008W	YGR186W	509	YGL048C	YGR184C
475	YDL150W	YPR187W	510	YPL042C	YPR187W

No	ID-1	ID-2	No	ID-1	ID-2
511	YCR040W	YDR404C	546	YBL008W	YDR308C
512	YER094C	YHR166C	547	YBL014C	YJR063W
513	YIL021W	YPL042C	548	YER177W	YPL242C
514	YCL031C	YJL148W	549	YMR236W	YOR174W
515	YJL140W	YKL058W	550	YGR005C	YJL056C
516	YGL025C	YOR210W	551	YPL122C	YPL139C
517	YGL025C	YOL005C	552	YDL017W	YJL173C
518	YBL084C	YDL007W	553	YDL164C	YLL004W
519	YDR404C	YPL042C	554	YOR038C	YOR194C
520	YBR058C	YDL147W	555	YGL201C	YOL115W
521	YDR167W	YOR210W	556	YML112W	YOR151C
522	YKL058W	YNL025C	557	YMR227C	YOL135C
523	YFL031W	YKR062W	558	YLR131C	YPR187W
524	YBR049C	YHR041C	559	YHR041C	YPR025C
525	YLR131C	YOR194C	560	YCR093W	YHR041C
526	YJL173C	YJL194W	561	YLL004W	YPR135W
527	YDL165W	YIL128W	562	YGL207W	YJL140W
528	YCR081W	YGR186W	563	YDL165W	YOR174W
529	YBL008W	YBR193C	564	YMR227C	YOR174W
530	YGR082W	YLR191W	565	YLR291C	YOR361C
531	YDR362C	YPR186C	566	YBR058C	YDL132W
532	YHR143WA	YPL042C	567	YBL008W	YOL005C
533	YBL008W	YPR187W	568	YGR005C	YJL127C
534	YJL001W	YMR275C	569	YGR186W	YHR084W
535	YER022W	YPL046C	570	YKR062W	YOR363C
536	YCL066W	YER148W	571	YCR081W	YDR308C
537	YBR193C	YDL165W	572	YBL084C	YLR263W
538	YBR193C	YJL056C	573	YDL164C	YOL115W
539	YOR224C	YPL042C	574	YLR274W	YNL290W
540	YMR229C	YNL216W	575	YKL139W	YOR174W
541	YDR059C	YER094C	576	YOR174W	YOR363C
542	YGL025C	YMR236W	577	YMR236W	YOR210W
543	YKL058W	YOR363C	578	YHR143WA	YOL051W
544	YFL031W	YPR187W	579	YBR088C	YML065W
545	YGR032W	YPL075W	580	YHR143WA	YOR038C

No	ID-1	ID-2	No	ID-1	ID-2
581	YDR118W	YIL156W	616	YNL290W	YPR019W
582	YBR198C	YOL051W	617	YGR013W	YKR086W
583	YCR081W	YPR168W	618	YFL031W	YHR143WA
584	YBR198C	YDL005C	619	YDL005C	YER171W
585	YOR124C	YOR261C	620	YGR005C	YHR041C
586	YDL097C	YNL172W	621	YDR146C	YKL028W
587	YGL043W	YJL006C	622	YJR112W	YLR263W
588	YCR040W	YGR005C	623	YOL051W	YOR224C
589	YBL008W	YOR210W	624	YGR005C	YLR131C
590	YMR309C	YPL237W	625	YHR013C	YPL175W
591	YOL135C	YPL046C	626	YIL156W	YOR261C
592	YDR427W	YKL213C	627	YDR092W	YFR050C
593	YMR227C	YNL236W	628	YDL140C	YPL046C
594	YDL007W	YJR099W	629	YHR118C	YOL094C
595	YDR211W	YOR361C	630	YER094C	YKL210W
596	YDR092W	YDR427W	631	YMR309C	YOR260W
597	YJR068W	YPR019W	632	YKL213C	YOR259C
598	YPR186C	YPR190C	633	YMR275C	YOR117W
599	YDL150W	YER148W	634	YJL197W	YOR362C
600	YDL005C	YGL035C	635	YER012W	YER144C
601	YBL093C	YIL021W	636	YGL025C	YIL021W
602	YBL093C	YOR210W	637	YGL112C	YHR058C
603	YDL007W	YMR275C	638	YER148W	YNL236W
604	YBR154C	YPL042C	639	YML112W	YPR086W
605	YBL093C	YBR154C	640	YFR004W	YGL240W
606	YMR275C	YOR259C	641	YCR018C	YNL216W
607	YER098W	YLR127C	642	YOR174W	YPL042C
608	YBR058C	YKL145W	643	YER133W	YIL162W
609	YDR054C	YPR135W	644	YHR143WA	YLR131C
610	YJL006C	YOL005C	645	YLR274W	YPR135W
611	YGR005C	YNL025C	646	YER171W	YGL025C
612	YDR211W	YER025W	647	YER017C	YIL046W
613	YDL126C	YGR253C	648	YGR119C	YJL050W
614	YKL028W	YOR151C	649	YHR058C	YIL128W
615	YCL066W	YDR404C	650	YHR058C	YMR227C

No	ID-1	ID-2	No	ID-1	ID-2
651	YCR042C	YOR210W	686	YER021W	YML111W
652	YDR311W	YGL025C	687	YKL028W	YOL135C
653	YLL039C	YPR108W	688	YDL140C	YOR174W
654	YGL201C	YJL090C	689	YML112W	YOR174W
655	YDL132W	YIL148W	690	YOL005C	YOL135C
656	YDL097C	YKL010C	691	YDR308C	YIR017C
657	YDL008W	YOR117W	692	YBL014C	YOR224C
658	YDL108W	YDR308C	693	YGR274C	YOL135C
659	YJR007W	YOL139C	694	YBL093C	YGL112C
660	YCR081W	YGR274C	695	YBR079C	YER025W
661	YGR184C	YPR108W	696	YCR042C	YDR404C
662	YIL021W	YLR131C	697	YDL007W	YDR059C
663	YDR156W	YJL025W	698	YBL093C	YKR062W
664	YIL143C	YOR174W	699	YER171W	YPL139C
665	YIL143C	YPL139C	700	YHR118C	YPR135W
666	YBL008W	YER022W	701	YBL084C	YER021W
667	YDL007W	YGL240W	702	YJL025W	YOR210W
668	YGL035C	YGR186W	703	YDL005C	YJL056C
669	YDR177W	YIL022W	704	YML043C	YOR224C
670	YBR193C	YHR084W	705	YER025W	YOR361C
671	YBR193C	YKL058W	706	YDR211W	YMR309C
672	YKR062W	YOR038C	707	YDR092W	YFR004W
673	YLR071C	YML112W	708	YJR094C	YKL089W
674	YKL058W	YML112W	709	YCR081W	YDL005C
675	YMR275C	YMR314W	710	YCR042C	YOR151C
676	YML112W	YOR224C	711	YDR108W	YDR182W
677	YGL043W	YHR058C	712	YGR186W	YNL236W
678	YJL127C	YOL005C	713	YGL240W	YPR108W
679	YBR058C	YDR427W	714	YGL048C	YKR094C
680	YDL005C	YPR056W	715	YGR274C	YOR337W
681	YER022W	YIR017C	716	YDR328C	YER098W
682	YLL039C	YOR117W	717	YDL165W	YER022W
683	YOR260W	YPL237W	718	YBL014C	YPR187W
684	YOL094C	YPR019W	719	YGL240W	YML092C
685	YGL025C	YKR062W	720	YCR093W	YGR104C

No	ID-1	ID-2	No	ID-1	ID-2
721	YOL038W	YOR249C	756	YDL020C	YOR249C
722	YJL056C	YKL028W	757	YGL025C	YGR186W
723	YGR253C	YKL213C	758	YFL039C	YKL013C
724	YML043C	YOR341W	759	YDR129C	YML064C
725	YER021W	YER100W	760	YDR207C	YDR311W
726	YDR394W	YER100W	761	YBR193C	YDR443C
727	YER021W	YMR275C	762	YBR038W	YDR477W
728	YML111W	YOR259C	763	YDR177W	YPL149W
729	YER094C	YER100W	764	YDR477W	YJL174W
730	YML043C	YOR340C	765	YDL097C	YLR102C
731	YKL145W	YKL213C	766	YOL005C	YPL046C
732	YGR005C	YPL042C	767	YER098W	YHR166C
733	YBR193C	YGR186W	768	YHR200W	YIL046W
734	YER025W	YOR260W	769	YBL084C	YDL097C
735	YJL127C	YOR224C	770	YMR275C	YPR103W
736	YDL097C	YFR036W	771	YFL031W	YOR224C
737	YKL139W	YOL051W	772	YBR193C	YIR018W
738	YMR275C	YOR362C	773	YER148W	YGL207W
739	YDR308C	YML098W	774	YDR087C	YJL148W
740	YIL021W	YOR174W	775	YDR443C	YOR174W
741	YHR065C	YNL216W	776	YKL028W	YLR131C
742	YAL009W	YNL126W	777	YKL213C	YOR362C
743	YBR193C	YML098W	778	YBR080C	YPR181C
744	YBR253W	YPR056W	779	YBR058C	YOR157C
745	YGR184C	YIL075C	780	YDL165W	YOR194C
746	YDL008W	YHR200W	781	YHR118C	YNL290W
747	YBR193C	YCR042C	782	YJL197W	YKL145W
748	YHR166C	YOR117W	783	YCR040W	YPR086W
749	YGL112C	YOR174W	784	YHR084W	YPR086W
750	YEL022W	YOR326W	785	YBL041W	YBR058C
751	YFR004W	YOR124C	786	YGL166W	YPR086W
752	YDR025W	YJL085W	787	YBL084C	YOR261C
753	YJL197W	YOR259C	788	YDR177W	YKL145W
754	YDL005C	YIL143C	789	YDL003W	YIL072W
755	YCL029C	YIL072W	790	YDR180W	YER132C

No	ID-1	ID-2	No	ID-1	ID-2
791	YGR274C	YNL236W	805	YDL126C	YFR052W
792	YBL008W	YDL140C	806	YDL008W	YHL027W
793	YBL008W	YBR154C	807	YGL171W	YJL148W
794	YKL010C	YOR117W	808	YMR260C	YOL139C
795	YCR042C	YHR143WA	809	YBR198C	YPR168W
796	YDR394W	YMR275C	810	YBR049C	YJL140W
797	YDR146C	YOL005C	811	YCL054W	YDR227W
798	YDL164C	YML065W	812	YDR201W	YDR488C
799	YOL135C	YPR056W	813	YGR030C	YMR059W
800	YJR094C	YOR058C	814	YBR193C	YGL207W
801	YGL207W	YGR186W	815	YEL032W	YNL088W
802	YER017C	YGL116W	816	YJL197W	YMR314W
803	YOL139C	YPL237W	817	YMR146C	YOR260W
804	YDR118W	YOR261C	818	YDL147W	YKL010C

Bibliography

Adamcsek, B., Palla, G., Farkas, I. J., Derenyi, I. and Vicsek, T.: 2006, CFinder: locating cliques and overlapping modules in biological networks, *Bioinformatics* **22**(8), 1021–1023.

Agrawal, R. and Srikant, R.: 1994, Fast algorithms for mining association rules in large databases., *VLDB*, pp. 487–499.

Aittokallio, T. and Schwikowski, B.: 2006, Graph-based methods for analysing networks in cell biology, *Brief Bioinform* **7**(3), 243–255.

Alashwal, H., Deris, S. and Othman, R. M.: 2006, One-class support vector machines for protein-protein interactions prediction, *International Journal of Biomedical Sciences* **1**(2), 120–127.

Ali, K. M. and Pazzani, M. J.: 1996, Error reduction through learning multiple descriptions., *Machine Learning* **24**(3), 173–202.

Alpaydin, E.: 2004, *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*, The MIT Press.

Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J. P., Chothia, C. and Murzin, A. G.: 2004, Scop database in 2004: refinements integrate structure and sequence family data, *Nucl. Acids Res.* **32**(suppl_1), D226–229.

Arnau, V., Mars, S. and Marn, I.: 2005, Iterative Cluster Analysis of Protein Interaction Data, *Bioinformatics* **21**(3), 364–378.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G.: 2000, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., *Nat Genet* **25**(1), 25–29.

Assenov, Y., Ramirez, F., Schelhorn, S.-E., Lengauer, T. and Albrecht, M.: 2008, Computing topological parameters of biological networks, *Bioinformatics* **24**(2), 282–284.

Bader, G. D., Donaldson, I., Wolting, C., Ouellette, B. F. F., Pawson, T. and Hogue, C. W. V.: 2001, BIND–The Biomolecular Interaction Network Database, *Nucl. Acids Res.* **29**(1), 242–245.

Bader, G. and Hogue, C.: 2003, An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics* **4**(1), 2.

Bader, J. S., Chaudhuri, A., Rothberg, J. M. and Chant, J.: 2004, Gaining confidence in high-throughput protein interaction networks., *Nature Biotechnology* **22**(1), 78–85.

Baldi, P. and Brunak, S.: 2001, *Bioinformatics: the machine learning approach*, MIT Press, Cambridge, MA, USA.

Bar-Joseph, Z., Gerber, G., Lee, T., Rinaldi, N., Yoo, J., Robert, F., Gordon, B., Fraenkel, E., Jaakkola, T., Young, R. and Gifford, D.: 21, Computational

discovery of gene modules and regulatory networks, *Nature Biotechnology* **21**(11), 1337–1342.

Barabasi, A.-L. and Oltvai, Z. N.: 2004, Network biology: understanding the cell's functional organization, *Nat Rev Genet* **5**(2), 101–113.

Bauer, E. and Kohavi, R.: 1999, An empirical comparison of voting classification algorithms: Bagging, boosting, and variants., *Machine Learning* **36**(1-2), 105–139.

Ben-David, S., Gehrke, J. and Schuller, R.: 2002, A theoretical framework for learning from a pool of disparate data sources, *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, New York, NY, USA, pp. 443–449.

Ben-Hur, A. and Noble, W. S.: 2005, Kernel methods for predicting protein-protein interactions, *Bioinformatics* **21**(suppl 1), i38–i46.

Ben-Hur, A. and Noble, W. S.: 2006, Choosing negative examples for the prediction of protein-protein interactions, *BMC Bioinformatics* **7**(Suppl 1), S2.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E.: 2000, The Protein Data Bank, *Nucl. Acids Res.* **28**(1), 235–242.

Bishop, C. M.: 1995, *Neural Networks for Pattern Recognition*, Oxford University Press.

Bishop, C. M.: 2006, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1 edn, Springer.

Bock, J. R. and Gough, D. A.: 2001, Predicting protein-protein interactions from primary structure, *Bioinformatics* **17**(5), 455–460.

Bradford, J. R. and Westhead, D. R.: 2004, Improved prediction of protein-protein binding sites using a support vector machines approach, *Bioinformatics* p. bti242.

Brohee, S. and van Helden, J.: 2006, Evaluation of clustering algorithms for protein-protein interaction networks, *BMC Bioinformatics* **7**(1), 488.

Browne, F., Wang, H., Zheng, H. and Azuaje, F.: 2006, An assessment of machine and statistical learning approaches to inferring networks of protein-protein interactions, *Journal of Integrative Bioinformatics* **3**(2), http://journal.imbio.de/index.php?paper_id=41.

Caragea, D., Silvescu, A. and Honavar, V.: 2004, A framework for learning from distributed data using sufficient statistics and its application to learning decision trees., *Int. J. Hybrid Intell. Syst.* **1**(2), 80–89.

Chakrabarti, D.: 2005, *Tools for Large Graph Mining*, PhD thesis, School of Computer Science, Canergie Mellon University.

Chawla, N. V., Japkowicz, N. and Kotcz, A.: 2004, Editorial: special issue on learning from imbalanced data sets., *SIGKDD Explorations* **6**(1), 1–6.

Chen, X.-W. and Liu, M.: 2005, Prediction of protein-protein interactions using random decision forest framework, *Bioinformatics* **21**(24), 4394–4400.

Cheng, J. and Baldi, P.: 2006, A machine learning information retrieval approach to protein fold recognition, *Bioinformatics* **22**(12), 1456–1463.

Cheng, J., Randall, A. Z., Sweredoski, M. J. and Baldi, P.: 2005, SCRATCH: a protein structure and structural feature prediction server, *Nucl. Acids Res.* **33**(suppl-2), W72–76.

Chia, J.-M. and Kolatkar, P.: 2004, Implications for domain fusion protein-protein interactions based on structural information, *BMC Bioinformatics* **5**(1), 161.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W.: 1998, A genome-wide transcriptional analysis of the mitotic cell cycle., *Mol Cell* **2**(1), 65–73.

Chothia, C. and Janin, J.: 1975, Principles of protein-protein recognition, *Nature* **256**, 705–708.

Chu, W., Ghahramani, Z., Krause, R. and Wild, D. L.: 2006, Identifying protein complexes in high-throughput protein interaction screens using an infinite latent feature model, *in* R. B. Altman, T. Murray, T. E. Klein, A. K. Dunker and L. Hunter (eds), *Pacific Symposium on Biocomputing*, World Scientific, pp. 231–242.

Davis, F. P. and Sali, A.: n.d., PIBASE: a comprehensive database of structurally defined protein interfaces, *Bioinformatics* **21**(9), 1901–1907.

De, S., Krishnadev, O., Srinivasan, N. and Rekha, N.: 2005, Interaction preferences across protein-protein interfaces of obligatory and non-obligatory components are different., *BMC Struct Biol* **5**.

Deng, M., Chen, T. and Sun, F.: 2003, An integrated probabilistic model for functional prediction of proteins, *RECOMB '03: Proceedings of the seventh annual international conference on Research in computational molecular biology*, ACM Press, New York, NY, USA, pp. 95–103.

Deng, M., Mehta, S., Sun, F. and Chen, T.: 2002, Inferring Domain-Domain Interactions From Protein-Protein Interactions, *Genome Res.* **12**(10), 1540–1548.

Dhamankar, R., Lee, Y., Doan, A., Halevy, A. Y. and Domingos, P.: 2004, imap: Discovering complex mappings between database schemas., *SIGMOD Conference*, pp. 383–394.

Dietterich, T. G.: 2000a, Ensemble methods in machine learning, *in* J. Kittler and F. Roli (eds), *Multiple Classifier Systems*, Vol. 1857 of *Lecture Notes in Computer Science*, Springer, pp. 1–15.

Dietterich, T. G.: 2000b, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization., *Machine Learning* **40**(2), 139–157.

Domingos, P.: 2000, A unified bias-variance decomposition for zero-one and squared loss, *AAAI–IAAI*, AAAI Press–The MIT Press, pp. 564–569.

Domingos, P.: 2003, Prospects and challenges for multi-relational data mining., *SIGKDD Explorations* **5**(1), 80–83.

Drummond, C. and Holte, R. C.: 2005, Learning to live with false alarms, *Workshop on Data Mining Methods for Anomaly Detection*, Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Duin, R.: 2002, The combining classifier: to train or not to train?, *Pattern Recognition, 2002. Proceedings. 16th International Conference on* **2**, 765–770 vol.2.

Dunn, R., Dudbridge, F. and Sanderson, C.: 2005, The use of edge-betweenness clustering to investigate biological function in protein interaction networks, *BMC Bioinformatics* **6**(1), 39.

Dzeroski, S.: 2003, Multi-relational data mining: an introduction., *SIGKDD Explorations* **5**(1), 1–16.

Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J. and Gerstein, M.: 2002, Bridging structural biology and genomics: assessing protein interaction data with known complexes, *Trends in Genetics* **18**(10), 529–536.

Espadaler, J., Romero-Isart, O., Jackson, R. M. and Oliva, B.: 2005, Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships, *Bioinformatics* **21**(16), 3360–3368.

Fauchere, J. L. and Pliska, V.: 1983, Hydrophobic parameters p of amino acid side chains from partitioning of n-acetyl-amino-acid amides, *J. Med.Chem.* **18**, 369–375.

Filkov, V. and Skiena, S.: 2004, Heterogeneous data integration with the consensus clustering formalism., *DILS*, pp. 110–123.

Fisher, R. A.: 1922, On the interpretation of χ^2 from contingency tables and the calculation of p, *Journal of the Royal Statistical Society* **85**(1), 87–94.

Fontana, P., Bindewald, E., Toppo, S., Velasco, R., Valle, G. and Tosatto, S. C. E.: 2005, The SSEA server for protein secondary structure alignment, *Bioinformatics* **21**(3), 393–395.

Friedman, N., Getoor, L., Koller, D. and Pfeffer, A.: 1999, Learning probabilistic relational models., *IJCAI*, pp. 1300–1309.

Futschik, M. E., Chaurasia, G. and Herzel, H.: 2007, Comparison of human protein protein interaction maps, *Bioinformatics* **23**(5).

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D. and Brown, P. O.: 2000, Genomic expression programs in the response of yeast cells to environmental changes, *Mol. Biol. Cell* **11**, 4241–4257.

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B. and Superti-Furga, G.: 2006, Proteome survey reveals modularity of the yeast cell machinery, *Nature* **440**, 631–636.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G.: 2002, Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature* **415**(6868), 141–7.

Getoor, L., Friedman, N., Koller, D. and Taskar, B.: 2001, Learning probabilistic models of relational structure., *ICML*, pp. 170–177.

Getoor, L., Friedman, N., Koller, D. and Taskar, B.: 2002, Learning probabilistic models of link structure., *Journal of Machine Learning Research* **3**, 679–707.

Gilchrist, M. A., Salter, L. A. and Wagner, A.: 2004, A statistical framework for combining and interpreting proteomic datasets, *Bioinformatics* **20**(5), 689–700.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., Mcdaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., Dasilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., Mckenna, M. P., Chant, J. and Rothberg, J. M.: 2003, A protein interaction map of drosophila melanogaster., *Science* **302**(5651), 1727–1736.

Goldberg, D. S. and Roth, F. P.: 2003, Assessing experimentally derived interactions in a small world, *Proceedings of the National Academy of Sciences of the United States of America* **100**(8), 4372–4376.

Gomez, S. M., Noble, W. S. and Rzhetsky, A.: 2003, Learning to predict protein-protein interactions from protein sequences, *Bioinformatics* **19**(15), 1875–1881.

Guharoy, M. and Chakrabarti, P.: 2007a, Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein protein interactions, *Bioinformatics* **23**(15), 1909–1918.

Guharoy, M. and Chakrabarti, P.: 2007b, Secondary structure based analysis and classification of biological interfaces: identification of binding motifs in protein protein interactions, *Bioinformatics* **23**(15), 1909–1918.

Gunasekaran, K., Tsai, C. J. and Nussinov, R.: 2004, Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers., *J Mol Biol* **341**(5), 1327–41.

Guo, Z., Li, Y., Gong, X., Yao, C., Ma, W., Wang, D., Li, Y., Zhu, J., Zhang, M., Yang, D. and Wang, J.: 2007, Edge-based scoring and searching method for identifying condition-responsive protein protein interaction sub-network, *Bioinformatics* **23**(16), 2121–2128.

Hall, M. A.: 1998, Correlation-based feature selection for machine learning, *Technical report*, University of Waikato.

Han, J.-D. J., Dupuy, D., Bertin, N., Cusick, M. E. and Vidal, M.: 2005, Effect of sampling on topology predictions of protein-protein interaction networks, *Nature Biotechnology* **23**(7), 839–844.

Hartemink, A. J. and Segal, E.: 2005, Session introduction: Joint learning from multiple types of genomic data., *Pacific Symposium on Biocomputing*.

Hastie, T., Tibshirani, R. and Friedman, J.: 2003, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer.

Hernandez, T. and Kambhampati, S.: 2004, Integration of biological sources: Current systems and challenges ahead., *SIGMOD Record* **33**(3), 51–60.

Ho, T. K.: 1998, The random subspace method for constructing decision forests, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**(8), 832–844.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M.,

Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Srensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D. and Tyers, M.: 2002, Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry, *Nature* **415**(6868), 180–183.

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R. S., Oughtred, R., Skrzypek, M. S., Weng, S., Wong, E. D., Zhu, K. K., Dolinski, K., Botstein, D. and Cherry, J. M.: 2008, Gene Ontology annotations at SGD: new data sources and annotation methods, *Nucl. Acids Res.* **36**(suppl1), D577–581.

Hoskins, J., Lovell, S. and Blundell, T. L.: 2006, An algorithm for predicting protein-protein interaction sites: Abnormally exposed amino acid residues and secondary structure elements, *Protein Sci* **15**(5), 1017–1029.

Huang, J. and Ling, C. X.: 2005, Using auc and accuracy in evaluating learning algorithms, *IEEE Trans. Knowl. Data Eng.* **17**(3), 299–310.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S. H.: 2000, Functional discovery via a compendium of expression profiles., *Cell* **102**(1), 109–126.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y.: 2001, A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci* **98**(8), 4569–4574.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M.: 2003, A bayesian networks approach for predicting protein-protein interactions from genomic data, *Science* **302**(5644), 449–453.

Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N.: 2001, Lethality and centrality in protein networks, *Nature* **411**(6833), 41–42.

Jones, S. and Thornton, J. M.: 1996, Principles of protein-protein interactions., *Proc Natl Acad Sci U S A* **93**(1), 13–20.

Jones, S. and Thornton, J. M.: 1997, Analysis of protein-protein interaction sites using surface patches., *J Mol Biol* **272**(1), 121–132.

Kabsch, W. and Sander, C.: 1983, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features., *Biopolymers* **22**(12), 2577–637.

Kanehisa, M. and Bork, P.: 2003, Bioinformatics in the post-sequence era, *Nature Genetics* **33**, 305 – 310.

Keedwell, E. and Narayanan, A.: 2005, *Intelligent Bioinformatics: The Application of Artificial Intelligence Techniques to Bioinformatics Problems*, Wiley.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L.,

Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R. and Hermjakob, H.: 2007, Intact–open source resource for molecular interaction data, *Nucl. Acids Res.* **35**(suppl-1), D561–565.

King, A. D., Przulj, N. and Jurisica, I.: 2004, Protein complex prediction via cost-based clustering, *Bioinformatics* **20**(17), 3013–3020.

Kohavi, R. and Wolpert, D.: 1996, Bias plus variance decomposition for zero-one loss functions, *ICML*, pp. 275–283.

Koike, A. and Takagi, T.: 2004, Prediction of protein-protein interaction sites using support vector machines, *Protein Eng. Des.* **17**.

Kuncheva, L. I.: 2004, *Combining Pattern Classifiers: Methods and Algorithms*, Wiley-Interscience.

Kuncheva, L. I. and Whitaker, C. J.: 2003, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Machine Learning* **51**(2), 181–207.

Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I. and Noble, W. S.: 2004, A statistical framework for genomic data fusion, *Bioinformatics* **20**(16), 2626–2635.

Lee, I., Date, S. V., Adai, A. T. and Marcotte, E. M.: 2004, A probabilistic functional network of yeast genes., *Science* **306**(5701), 1555–1558.

Li, D., Li, J., Ouyang, S., Wang, J., Wu, S., Wan, P., Zhu, Y., Xu, X. and He, F.: 2006, Protein interaction networks of saccharomyces cerevisiae, caenorhabditis elegans and drosophila melanogaster: large-scale organization and robustness., *Proteomics* **6**(2), 456–461.

Li, M.-H., Lin, L., Wang, X.-L. and Liu, T.: 2007, Protein protein interaction site prediction based on conditional random fields, *Bioinformatics* **23**(5), 597–604.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E. and Vidal, M.: 2004, A map of the interactome network of the metazoan c. elegans., *Science* **303**(5657), 540–543.

Li, W., Han, J. and Pei, J.: 2001, Cmar: Accurate and efficient classification based on multiple class-association rules., *in* N. Cercone, T. Y. Lin and X. Wu (eds), *ICDM*, IEEE Computer Society, pp. 369–376.

Lin, N., Wu, B., Jansen, R., Gerstein, M. and Zhao, H.: 2004, Information assessment on predicting protein-protein interactions, *BMC Bioinformatics* **5**(1), 154.

Liu, B., Hsu, W. and Ma, Y.: 1998, Integrating classification and association rule mining., *KDD*, pp. 80–86.

Livingstone, C. D. and Barton, G. J.: 1993, Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation., *Computer Applications in the Biosciences* **9**(6), 745–756.

Lu, L. J., Xia, Y., Paccanaro, A., Yu, H. and Gerstein, M.: 2005, Assessing the limits of genomic data integration for predicting protein networks, *Genome Res.* **15**(7), 945–953.

Merugu, S. and Ghosh, J.: 2005, A distributed learning framework for heterogeneous data sources, *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM Press, New York, NY, USA, pp. 208–217.

Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B.: 2002, Mips: a database for genomes and protein sequences, *Nucl. Acids Res.* **30**(1), 31–34.

Mintseris, J. and Weng, Z.: 2005, Structure, function, and evolution of transient and obligate protein-protein interactions., *Proc Natl Acad Sci U S A* .

Mitchell, T. M.: 1997, *Machine Learning*, McGraw-Hill, New York.

Mjolsness, E. and Decoste, D.: 2001, Machine learning for science: State of the art and future prospects, *Science* **293**(5537), 2051–2055.

Narayanan, A., Keedwell, E. and Olsson, B.: 2002, Artificial intelligence techniques for bioinformatics, *Applied Bioinformatics* **1**(4), 191–222.

Neuvirth, H., Raz, R. and Schreiber, G.: 2004, Promate: a structure based prediction program to identify the location of protein-protein binding sites., *J Mol Biol* **338**(1), 181–199.

Neuvirth, H., Raz, R. and Schreiber, G.: n.d., Promate: a structure based prediction program to identify the location of protein-protein binding sites., *J Mol Biol* (1), 181–199.

Nooren, I. M. and Thornton, J. M.: 2003, New embo member's review: Diversity of protein-protein interactions, *EMBO J.* **22**(14), 3846–3492.

Park, S., Reyes, J., Gilbert, D., Kim, J. and Kim, S.: 2009, Prediction of protein-protein interaction types using association rule based classification, *BMC Bioinformatics* **10**(1), 36.

Pereira-Leal, J. B., Enright, A. J. and Ouzounis, C. A.: 2004, Detection of functional modules from protein interaction networks., *Proteins* **54**(1), 49–57.

Phizicky, E. and Fields, S.: 1995, Protein-protein interactions: methods for detection and analysis, *Microbiol. Rev.* **59**(1), 94–123.

Qi, Y., Balem, F., Faloutsos, C., Klein-Seetharaman, J. and Bar-Joseph, Z.: 2008, Protein complex identification by supervised graph local clustering, *Bioinformatics* **24**(13), i250–268.

Qi, Y., Bar-Joseph, Z. and Klein-Seetharaman, J.: 2006, Evaluation of different biological data and computational classification methods for use in protein interaction prediction, *Proteins: Structure, Function, and Bioinformatics* **63**(3), 490–500.

Qi, Y., Klein-Seetharaman, J. and Bar-Joseph, Z.: 2005, Random forest similarity for protein-protein interaction prediction from multiple sources., *Pacific Symposium on Biocomputing*, World Scientific.

Quinlan, J. R.: 1993, *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rahm, E. and Bernstein, P. A.: 2001, A survey of approaches to automatic schema matching., *VLDB J.* **10**(4), 334–350.

Ravasz, E. and Barabasi, A. L.: 2003, Hierarchical organization in complex networks, *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)* **67**(2).

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L.: 2002, Hierarchical organization of modularity in metabolic networks, *Science* **297**(5586), 1551–1555.

Reguly, T., Breitkreutz, A., Boucher, L., Breitkreutz, B.-J., Hon, G., Myers, C., Parsons, A., Friesen, H., Oughtred, R., Tong, A., Stark, C., Ho, Y., Botstein, D., Andrews, B., Boone, C., Troyanskya, O., Ideker, T., Dolinski, K., Batada, N. and Tyers, M.: 2006, Comprehensive curation and analysis of global interaction networks in saccharomyces cerevisiae, *Journal of Biology* **5**(4), 11.

Reinoso, J., Silvescu, A., Caragea, D., Pathak, J. and Honavar, V.: 2003, Information extraction and integration from heterogeneous, distributed, autonomous information sources : A federated ontology-driven query-centric approach., *IRI*, pp. 183–191.

Reyes, J. A. and Gilbert, D.: 2007, Prediction of protein-protein interactions using one-class classification methods and integrating diverse data, *Journal of Integrative Bioinformatics* **4**(3), http://journal.imbio.de/index.php?paper_id=77.

Reyes, J. A. and Gilbert, D.: 2008, Combining one-class classification models based on diverse biological data for prediction of protein-protein interactions, *in* A. Bairoch, S. C. Boulakia and C. Froidevaux (eds), *DILS*, Vol. 5109 of *Lecture Notes in Computer Science*, Springer, pp. 177–191.

Rives, A. W. and Galitski, T.: 2003, Modular organization of cellular networks, *Proceedings of the National Academy of Sciences of the United States of America* **100**(3), 1128–1133.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P. and Vidal, M.: 2005, Towards a proteome-scale map of the human proteinprotein interaction network, *Nature* .

Sanner, M. F., Olson, A. J. and Spehner, J. C.: 1996, Reduced surface: An efficient way to compute molecular surfaces, *Biopolymers* **38**(3), 35–320.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. and Williamson, R. C.: 2001, Estimating the support of a high-dimensional distribution, *Neural Computation* **13**(7), 1443–1471.

Scholtens, D., Vidal, M. and Gentleman, R.: 2005, Local modeling of global interactome networks, *Bioinformatics* **21**(17), 3548–3557.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T.: 2003, Cytoscape: a software environment for integrated models of biomolecular interaction networks., *Genome Research* **13**(11), 2498–2504.

Sharan, R., Ideker, T., Kelley, B., Shamir, R. and Karp, R. M.: 2005, Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data., *Journal of Computational Biology* **12**(6), 835–846.

Shoemaker, B. A. and Panchenko, A. R.: 2007a, Deciphering proteinprotein interactions. part i. experimental techniques and databases, *PLoS Comput Biol* **3**.

Shoemaker, B. A. and Panchenko, A. R.: 2007b, Deciphering proteinprotein interactions. part ii. computational methods to predict protein and domain interactions partners, *PLoS Comput Biol* **3**.

Spirin, V. and Mirny, L. A.: 2003, Protein complexes and functional modules in molecular networks, *Proceedings of the National Academy of Sciences of the United States of America* **100**(21), 12123–12128.

Stein, L.: 2003, Integrating biological databases, *Nat Rev Genet* **4**(5), 337–345.

Szilágyi, A., Grimm, V., Arakaki, A. K. and Skolnick, J.: 2005, Prediction of physical protein-protein interactions, *Physical Biology* **2**(2), S1–S16.

Tang, E. K., Suganthan, P. N. and Yao, X.: 2006, An analysis of diversity measures, *Machine Learning* **65**(1), 247–271.

Tax, D. M. J.: 2001, *One-class classification.*, PhD thesis, Delft University of Technology.

Tax, D. M. J. and Duin, R. P. W.: 2004, Support vector data description., *Machine Learning* **54**(1), 45–66.

Tsymbal, A., Pechenizkiy, M. and Cunningham, P.: 2005, Diversity in search strategies for ensemble feature selection, *Information Fusion* **6**(1), 83–98.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M.: 2000, A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae, *Nature* **403**(6770), 623–627.

Vakser, I. A.: 2004, Protein-protein interfaces are special, *Structure* **12**, 910–912.

Van Berlo, R. J. P., Wessels, L. F., Ridder, D. D. E. and Reinders, M. J. T.: 2007, Protein complex prediction using an integrative bioinformatics approach, *J Bioinform Comput Biol* **5**(4), 839–864.

Vapnik, V. N.: 1998, *Statistical Learning Theory*, John Willey and Sons.

Varma, S. and Simon, R.: 2006, Bias in error estimation when using cross-validation for model selection, *BMC Bioinformatics* **7**(1), 91.

Viksna, J., Gilbert, D. and Torrance, G. M.: 2003, Protein structure comparison based o n profiles of topological motifs: a feasible way to deal with information from negative examples, *German Conference on Bioinformatics*, pp. 159–165.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A. and Bork, P.: 2005, STRING: known and predicted protein-protein associations, integrated and transferred across organisms, *Nucl. Acids Res.* **33**(suppl_1), D433–437.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Bork, P.: 2002, Comparative assessment of large-scale data sets of protein-protein interactions., *Nature* **417**(6887), 399–403.

Wang, H., Segal, E., Ben-Hur, A., Koller, D. and Brutlag, D. L.: 2005, Identifying protein-protein interaction sites on a genome-wide scale, *in* L. K. Saul, Y. Weiss and L. Bottou (eds), *Advances in Neural Information Processing Systems 17*, MIT Press, Cambridge, MA, pp. 1465–1472.

Wang, H., Segal, E., Ben-Hur, A., Li, Q.-R., Vidal, M. and Koller, D.: 2007, Insite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale, *Genome Biology* **8**(9), R192.

Wilcoxon, F.: 1945, Individual comparisons by ranking methods, *Biometrics Bulletin* **1**(6), 80–83.

William S. J. Valdar, J. M. T.: 2001, Protein-protein interfaces: Analysis of amino acid conservation in homodimers, *Proteins: Structure, Function, and Genetics* **42**(1), 108–124.

Witten, I. H. and Frank, E.: 2005, *Data Mining: Practical machine learning tools and techniques*, 2nd edn, Morgan Kaufmann, San Francisco.

Wojcik, J. and Schachter, V.: 2001, Protein-protein interaction map inference using interacting domain profile pairs, *Bioinformatics* **17**, S296–305.

Wolkenhauer, O. and Gierl, L.: 2003, From bioinformatics to systems biology: Towards an understanding of the dynamics of life, *Traditio Et Innovatio* **8**(2), 10–13.

Wong, L.: 2002, Technologies for integrating biological data., *Briefings in Bioinformatics* **3**(4), 389–404.

Wu, S. and Zhang, Y.: 2008, A comprehensive assessment of sequence-based and template-based methods for protein contact prediction, *Bioinformatics* **24**(7), 924–931.

Yamanishi, Y., Vert, J.-P. and Kanehisa, M.: 2004, Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics* **20**(suppl1), i363–370.

Yin, X., Han, J., Yang, J. and Yu, P. S.: 2004, Crossmine: Efficient classification across multiple database relations., *ICDE*, pp. 399–411.

Yoo, P., Sikder, A., Zhou, B. and Zomaya, A.: 2008, Improved general regression network for protein domain boundary prediction, *BMC Bioinformatics* **9**(Suppl 1).

Yook, S. H., Oltvai, Z. N. and Barabási, A. L.: 2004, Functional and topological characterization of protein interaction networks., *Proteomics* **4**(4), 928–942.

Yule, G. U.: 1900, On the association of attributes in statistics, *Philosophical Transactions of the Royal Society of London* **A**(194), 257–319.

Zhang, L., Wong, S., King, O. and Roth, F.: 2004, Predicting co-complexed protein pairs using genomic and proteomic data integration, *BMC Bioinformatics* **5**(1), 38.

Zhang, S.-H., Ning, X.-M., Liu, H.-W. and Zhang, X.-S.: 2006, Prediction of protein complexes based on protein interaction data and functional annotation data using kernel methods, *in* D.-S. Huang, K. Li and G. W. Irwin (eds), *ICIC (3)*, Vol. 4115 of *Lecture Notes in Computer Science*, Springer, pp. 514–524.

Zhou, H.-X. and Qin, S.: 2007, Interaction-site prediction for protein complexes: a critical assessment, *Bioinformatics* **23**(17), 2203–2209.

Zhu, H., Domingues, F. S., Sommer, I. and Lengauer, T.: 2006, Noxclass: prediction of protein-protein interaction types, *BMC Bioinformatics* **7**(27), 1–15.